Xenacoelomorpha: The "simple" key to bilaterian ancestry?

Steven Müller

Department of Genetics, Evolution and Environment
University College London
Primary supervisor: Prof. Maximilian J. Telford
Secondary supervisor: Prof. Christophe Dessimoz

Submitted for the Degree of Doctor of Philosophy

London, January 2019

Declaration

I, Steven Müller, hereby confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Xenacoelomorpha (comprising Xenoturbellida, Acoela and Nemertodermatida) is a clade of marine worms whose position in the tree of life is still in debate. Several phylogenetic analyses have shown them to be placed at the base of all bilaterian animals (e. g. chordates, arthropods) or at a more derived position as sister group to the Ambulacraria (echinoderms and hemichordates) within the Bilateria. A key characteristic is the absence of traits found in other bilaterian animals.

Orthogroups are groups of orthologous genes found in several organisms. Orthologues are assumed to retain the same function. These functions would be specific to the clade where an orthogroup is prevalent. I investigate a method to automatically establish and validate orthogroups specific to Bilateria, Protostomia and Deuterostomia. These genes could be relevant for the clades' respective emergence and differences. These sets will also help to ascertain what genes/functions are absent from Xenacoelomorpha.

MicroRNAs (miRNAs) are small non-coding RNA molecules involved in RNA silencing and post-transcriptional regulation of gene expression. MiRNAs have not been extensively studied in the Xenaceolomorpha. I introduce a fully automatic miRNA detection pipeline to infer and confirm the existence of pre-miRNA sequences in the genome of *Xenoturbella bocki* as well as predict miRNA candidates from several xenacoel genomes. I report previously undetected miRNA families and opine that previous analyses on Acoelomorpha failed due to loss caused by the higher evolutionary rate when compared to the Xenoturbellida.

Statement of Scientific Impact

The study of evolution requires understanding of how species and genes are related. It is fascinating to understand, how millions of species can originate from a common ancestor. The study of these relationships is not just important to satisfy our own curiosity about how we came to be, but also about how knowledge from other species can benefit us in how we will be.

Xenacoelomorpha represent a case that challenges our understanding of how species are related and how this is reflected by their physiology. These marine worms are capable to survive without the need for specialised organs or an excretory system. Seen by some as an early offshoot of Bilateria, others consider them to be more derived, i.e. having evolved from an arguably more complex ancestor. Both cases imply a lack of many features that other organisms need for survival. Studying these marine organisms can show us how a lack of otherwise common characters can be feature rather than a bug.

The results of my investigations in Xenacoelomorpha will further the discussion about arguably absent genetic characters. The methods described here could be applied to other species of interest to find candidates for supposedly absent genetic traits. The widespread application of improved methodology will allow us to further our knowledge about investigated organisms, but also show us the limits of our current methods.

In my projects I developed several methods that aim to decrease the manual handling of data. The application of automated pipelines has several advantages. The reduction of *ad hoc* approaches leads to a higher standardisation of experiments. This increases our ability to both replicate and compare results. Automation also allows for a higher throughput of data resulting in a faster accumulation of new knowledge.

However, automation must be rigorously checked and constantly re-evaluated. Throughout my projects I encountered many obstacles that increased the difficulty of applying automated methods. Data quality and biases as well as inherent methodological drawbacks require careful consideration about applying specific approaches when facing non-trivial problems. With the examples examined here, I hope to describe common mistakes and how we as a community can avoid them.

Acknowledgements

I would like to thank my supervisors, Max Telford and Christophe Dessimoz, for their supervision and professional guidance throughout my PhD. Both of my supervisors helped me to expose myself to new challenges. I thank Max Telford for giving me a chance to apply my skills and experience to the field of evolutionary biology, a field I was not very familiar with before starting my PhD. Max' zoological background was an invaluable input to the development and discussion of new ideas. I thank Christophe Dessimoz for providing complementary guidance in the field of Bioinformatics. The successful application of my ideas was only possible through the combination of both biological and computational concepts. I would also like to thank Paola Oliveri for her comments and critical input regarding my upgrade report, group meeting presentations and in general.

I want to thank all my colleagues that I was lucky enough to share my time with during this PhD. From the Telford, Dessimoz, Yang and Oliveri groups, I would like to thank Anne, Fraser, Helen, Irepan, Johannes, Laura, Paschalia, Philipp, Ivana, Karina, Xiyun, Ania, David, Libero and Natalie. I would like to extend my thanks to other colleagues in and outside of UCL, including Alun, Çağrı, Daniel, Fabian, Fritz, Jan, Prudence and Víctor. I am grateful not just for the professional advice and support, but also the friendships we have made along the way. Without this network I would have no doubt struggled even more than I already had. Each and everyone of you has provided me with insight and knowledge in my professional and personal development. All the comments, discussions, jokes, games and experiences we shared made this whole endeavour worthwhile.

Abschließend möchte ich mich ganz herzlich bei meiner Familie für ihre anhaltende Unterstützung bedanken: Meinen Eltern, meinen Schwestern, Oma Elli und ganz besonders meiner Frau Hai. Vielen Dank, dass ihr für mich da seid, nicht nur während der letzten 5 Jahre. Es ist wichtig jemanden zu haben, die einem sowohl in guten wie in schlechten Zeiten Rückhalt bieten.

Contents

List of Figures

1. Introduction to Xenacoelomorpha

The phylum Xenacoelomorpha is a proposed clade of wormlike marine worms [Philippe et al., 2011]. It comprises the Xenoturbellida (6-7 described species, Rouse et al. [2016], Nakano et al. [2017]), Acoela (20 families, close to 400 species) and Nemertodermatida (10 species).

Xenacoelomorphs can be found in diverse marine environments across the whole globe. Most of the species described are free living, but some have been found to live as parasites or endosymbionts within other marine species (e.g. *Meara stichopi* which lives in the gut of the sea cucumber *Parastichopus tremulus*). They display a wide spectrum of colouration and a varying degree of pigmentation. Their body shapes range from a compact oval shape to long and slender (fig. 1.1).

Their body plan is bilaterally symmetrical and arguably simple. This observed simplicity has raised questions about their relationship with other animals. A debate that is still going on to this day (fig. 1.2).

1.1. Acoelomorpha

1.1.1. Morphology

Like other bilaterians, acoelomorphs have three germ layers and a body plan with an anterior-posterior as well as dorsal-ventral body axes making them bilaterally symmetrical. But unlike most other bilaterians acoelomorphs possess very few distinct structures. They do not have coelomic cavities, organs or a circulatory system. Instead of a through gut with separate mouth and anus, they only have a ventral mouth opening that connects to a sac-like gut. A shared structure amongst Acoelomorpha is the existence of

Figure 1.1.: Diversity of the Xenacoelomorpha. Top row: lateral and dorsal view of *Xenoturbella bocki* (images courtesy of A.-C. Zakrzewski) - bottom row clockwise from left: *Meara stichopi*, *Paratomella rubra*, *Symsagittifera roscoffensis* (Telford group).

a statocyst, a spherical sensory receptor containing a mineralised mass (statolith) and sensory hairs to detect orientation and acceleration [Achatz et al., 2013].

The acoelomorph's acoelomate nature and lack of a digestive tract lead to their original classification as members of the phylum Platyhelminthes, a clade of flat worms that show a similar simplicity in body organisation. More specifically, they were linked to the Turbellaria, platyhelminths that are free-living, and not the Neodermata, platyhelminths that are exclusively parasites and descended from within the Turbellaria. Platyhelminthes were considered by many to be the sister to all other bilaterians based on their simplicity and lack of more complex characters. Together with the acoelomorphs they were thought to represent an intermediate off-shoot between the ancestor of Eumetazoa (Cnidaria + Bilateria) and the ancestor of Bilateria.

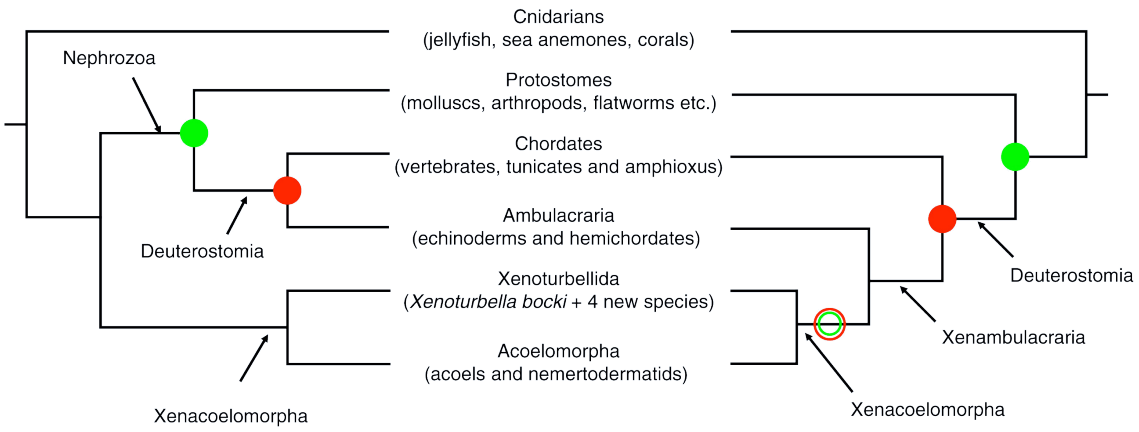


Figure 1.2.: Modified from Telford and Copley [2016]: Phylogenetic trees showing alternative hypotheses for placement of Xenacoelomorpha - left: xenacoelomorphs diverging before the appearance of bilaterian (green circle) and deuterostome characters (red circle), right: xenacoelomorphs are sister to Ambulacraria implying the loss of bilaterian and deuterostome characters in the xenacoelomorph ancestor (empty red+green circle).

1.1.2. Molecular analyses

Phylogenetic analyses

Acoelomorphs were initially grouped with the Platyhelminthes based on morphological characters. The availability of molecular analyses would allow to further resolve the relationship within the group.

18S ribosomal DNA sequences (18S rDNA, coding for a component of the small eukaryotic ribosomal subunit) was the first phylogenetic marker used in several studies to assess the phylogenetic position of Aceolomorpha. Katayama et al. [1996] sequenced almost complete 18S rRNA sequences for aceols and turbellarians. They reconstructed phylogenetic trees adding 18S rDNA sequences from several diploblasts and yeast, but did not include any other bilaterian species. They found acoels and turbellarians to form a monophyletic clade, with Acoela placed as sister to all remaining Turbellaria.

A major change of the acoelomorphs' phylogenetic relations was revealed when Platyhelminthes were found not to be sister to all other bilaterians. Using 18S rDNA several studies concluded that Platyhelminthes belonged to the Protostomia inside the Bilateria [Balavoine, 1997, Carranza et al., 1997], deriving from within the Lophotrochozoa.

This drastically changes the interpretation of the Platyhelminthes' simple morphology from an ancestral state to a derived state involving secondary loss of characters. Campos et al. [1998] found 2 acoel species to group with Tricladida/Seriata supporting the monophyly of Platyhelminthes and Acoela. Contrary, Ruiz-Trillo et al. [1999] found the grouping to be paraphyletic. While Platyhelminthes retained their positon amongst Protostomia, Acoela were placed outside the Bilateria branching first after the split from Cnidaria. Littlewood et al. [1999] also found acoels to be outside the Bilateria, but referred to the estimated long branches leading to Acoela as problematic in that Long Branch Attraction, a systematic error causing distantly related species to appear closely related, could be responsible for this result obscuring the true phylogenetic signal.

More studies confirmed the acoels' position as sister to all other bilaterians, but faced issues inferring a monophyletic clade for acoels and nemertodermatids. Jondelius et al. [2002] used 18S rDNA, COI and cytochrome b (Cytb) sequences, but not all reconstruction methods resulted in a monophyletic grouping of Acoelomorpha. Ruiz-Trillo et al. [2002] used myosin II gene sequences that supported a monophyletic grouping which was, however, disrupted when they tried to verify the results using 18S rDNA data. Telford et al. [2003] also inferred paraphyletic Acoelomorpha, but found that a monophyletic scenario was not significantly worse in comparison.

Studies of specific molecular characters

The idea of Acoelomorpha representing an intermediary branch between diploblasts and triploblasts was further supported by the lack of certain other molecular characters found in other Bilateria.

Hox genes are are an important class of genes that regulate embryonic development and body patterning. They share a characteristic binding motif encoded by the homeobox ("Hox") and have been associated with the patterning of the anterior-posterior axis in the majority of metazoans. Their link to specific body segments implies their association with body plan complexity. The bilaterian ancestor had at least 7 Hox genes [de Rosa et al., 1999] and 3 ParaHox genes (inferred to be from an ancient duplication of the Hox cluster, Brooke et al. [1998]) while cnidarians only possess 2 Hox and 2 ParaHox genes.

The lack of body segmentation or patterning in Acoelomorpha raises the question if this could be reflective of their Hox gene complement. Cook et al. [2004] searched for

Hox genes in the acoels *S. roscoffensis* and *P. rubra* and found representatives of 3 out of 7 Hox genes and 1 out of 3 ParaHox genes present in Bilateria. Jiménez-Guri et al. [2006] sequenced a representative of the previously missing gene group (ParaHox Xlox) in the nemertodermatid *Nemertodermatida westbladi*, but also did not find evidence for the bilaterian expansion in the anterior and central Hox gene groups (fig. 1.3).

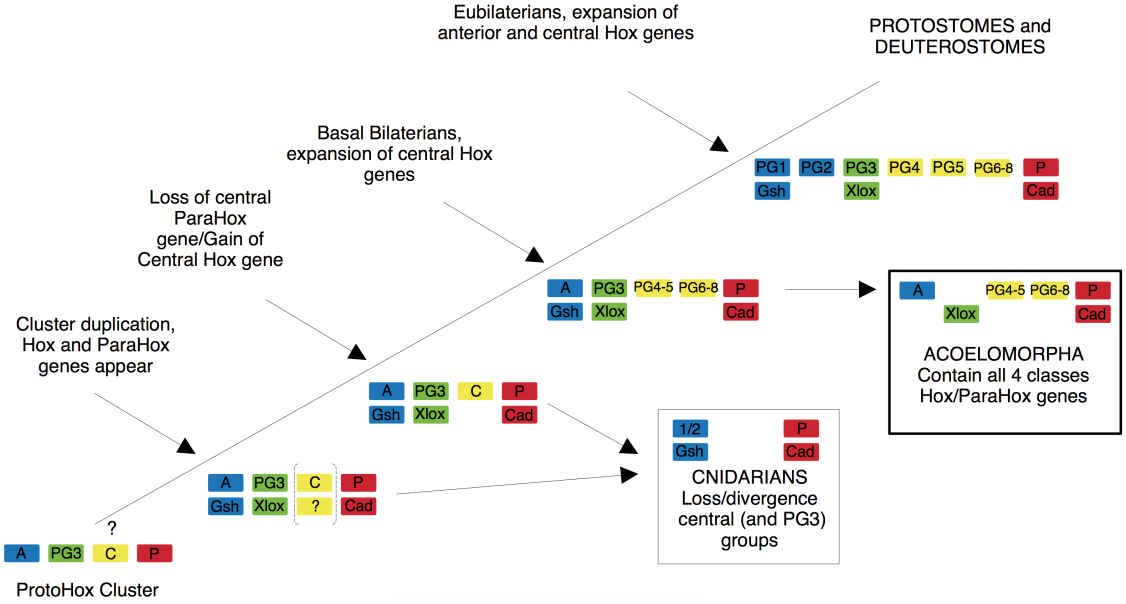


Figure 1.3.: Evolution of Hox and ParaHox gene groups proposed by Jiménez-Guri et al. [2006]. Hox genes are involved in the anterior-posterior patterning during embryonal development. Acoelomorphs are missing bilaterian expansion of anterior and central Hox gene groups found in protostomes and deuterostomes.

Another molecular character shared across Bilateria, but said to absent in acoelomorphs are microRNAs. MicroRNAs (miRNAs) are small non-coding RNA sequences involved in RNA silencing. They are important in post-transcriptional regulation of gene expression and their high conservation between distantly related species implies a restricted rate of evolution. *let-7* is one of the most well conserved miRNAs in Bilateria, but Pasquinelli et al. [2000] were unable to detect evidence for it in 3 acoel species. Sempere et al. [2007] extended the search to other miRNAs which they found to be conserved across Bilateria, but could only find evidence for 6 out of the 16 miRNAs tested (fig. 1.4).

The absence of certain morphological as well as molecular characters which are other-

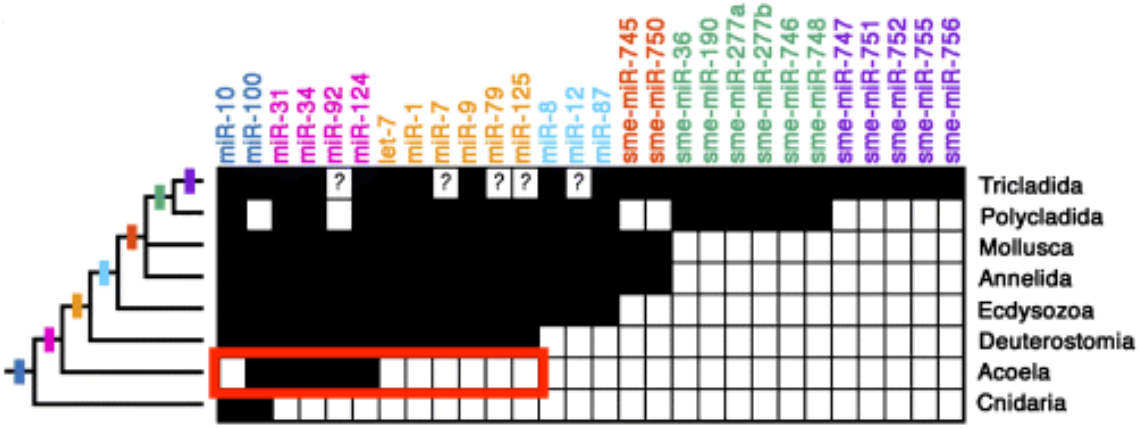


Figure 1.4.: MicroRNAs across Eumetazoa as sequenced by Sempere et al. [2007]. MicroRNAs are involved in the regulation of gene expression. Emergence of new microRNAs has been associated with more complex tissues and structures. Red box - Acoela lack several microRNAs conserved across Protostomia and Deuterostomia due to lack of microRNA evidence sequenced from *S. roscoffensis*.

wise present across the Bilateria seem to make a strong case for the intermediary position of Acoelomorpha. If this is their true phylogenetic position then the Acoelomorpha would represent an important taxon to study if we wish to explore the transition between the eumetazoan and bilaterian ancestors. The simple morphology and position of acoelomorphs inform a much simpler ancestor of all bilaterian animals than the ancestor based on shared protostome and deuterostome features (fig. 1.5). Investigations of another simple looking marine worm, *Xenoturbella bocki*, however, spawned new hypotheses about the acoelomorphs' position amongst Bilateria.

1.2. Xenoturbellida

1.2.1. Morphology

Xenoturbella bocki is another marine worm showing a similar lack of distinct characters like Acoelomorpha. Akin to acoelomorphs it was originally classified alongside the platyhelminths. It was discovered off the Swedish west coast by Swedish zoologist Sixten Bock in 1915 and first described by Einar Westblad in 1949 [Westblad, 1949] using a

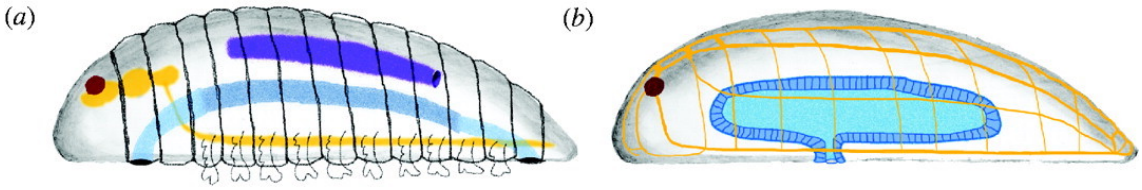


Figure 1.5.: Modified from Hejnol and Martindale [2008]: Hypotheses for the ancestor of protostomes and deuterostomes (a) and the ancestor of all Bilateria (b) when placing Xenacoelomorpha as sister to all other bilaterians. The inferred urbilaterian lacks features such as body segmentation (black segments), heart (dark blue, dorsal side) or appendages (ventral markings), has a blind-gut (light blue) and a less condensed nervous system (yellow).

collection of Bock's sections, drawings, photographs and notes.

Bock's notes describe *X. bocki*'s morphology as extremely simple, stating that the found specimen resemble slimy lumps more than living animals. He compared the absence of complex structures in *X. bocki* with the similar appearance of Acoelomorpha. Bock's notes emphasise the phylogenetic importance of this newly discovered species, but he remained uncertain about its relationship to other animals.

Bock's notes describe only one character that defies the simple morphology: *X. bocki*'s body wall is described as "remarkably similar to that of the Enteropneusta [hemichordate acorn worms]". However, the observed lack (e.g. blood vessels) or "primitive state" (e.g. reproductive organs) of features led to the conclusion that *X. bocki* represents a primitive organism ("lowest level of Bilateria") and should be grouped with the Platyhelminthes alongside Acoelomorpha.

Through his own observations, Westblad made a first direct connection of *X. bocki* to Acoelomorpha. The system of epithelial muscle fibres in *Xenoturbella* is similar to that of *Nemertoderma* (Nemertodermatida). Based on the overall lack of characters and its potential affiliation with the Turbellaria (free living Platyhelminthes) he proposed the name *Xenoturbella* (*xeno* for "strange" + *turbella*).

Franzén and Afzelius [1987] strengthened the connection to Acoelomorpha when examining *Xenoturbella*'s ciliated epidermis. The cilium contains the typical 9 + 2 pattern of microtubule doublets, but doublets 4 - 7 end before the tip of the cilium resulting in a shelf below the tip similar to what has been described for Nemertodermatida and lower Acoela [Tyler, 1979], indicating a close relationship between these three clades.

This shelf like structure is not unique to *X. bocki* and Acoelomorpha, as it has also been observed in *Glossobalanus minutus* (phylum Hemichordata, Pardos [1988], fig. 1.6).

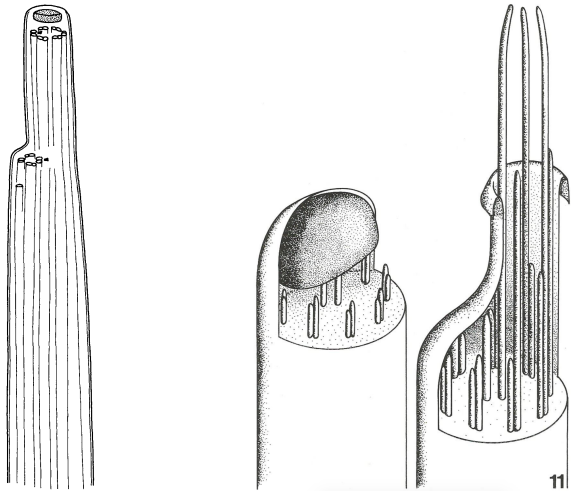


Figure 1.6.: Cilia comparison between *Xenoturbella* (left, drawing from Franzén and Afzelius, 1987) and type I and type II cilia of hemichordate *Glossobalanus minutus* (right, drawing from Pardos, 1988). Cilia of *X. bocki* share the electron dense distal cap with type I cilia of *G. minutus* and the shelf like structure of microtubules ending before the tip with type II cilia of *G. minutus*.

Over the following years, more morphological evidence to group *Xenoturbella* and Acoelomorpha into one clade were found: Franzén and Afzelius [1987], Rohde et al. [1988] and Lundin and Hendelberg [1998] all observed many similar characteristics when they compared the epidermal ciliary structure of *Xenoturbella* and acoelomorphs. These characteristics include the shelf towards the distal end of the cilium and a cup-shaped structure at the proximal end of the cilium (fig. 1.6). Lundin listed 9 possible synapomorphic characters regarding ciliary structure [1998] and found the same process of epidermal degeneration [Lundin, 2001] for *Xenoturbella* and acoelomorphs.

Achatz et al. [2013] questioned the usefulness of ciliary structures as phylogenetic characters. They argued that these morphological similarities could either be shared plesiomorphies (remnants of a common ancestor) or convergent adaptions caused by a similar marine lifestyle.

Questions about the turbellarian positioning of *X. bocki* were raised by Franzén (Fran-

zén [1956]; quoted in Franzén and Afzelius [1987]) after finding that *Xenoturbella* would be the only species amongst the Turbellaria to have retained a "primitive" type of spermatozoon (fig. 1.7).

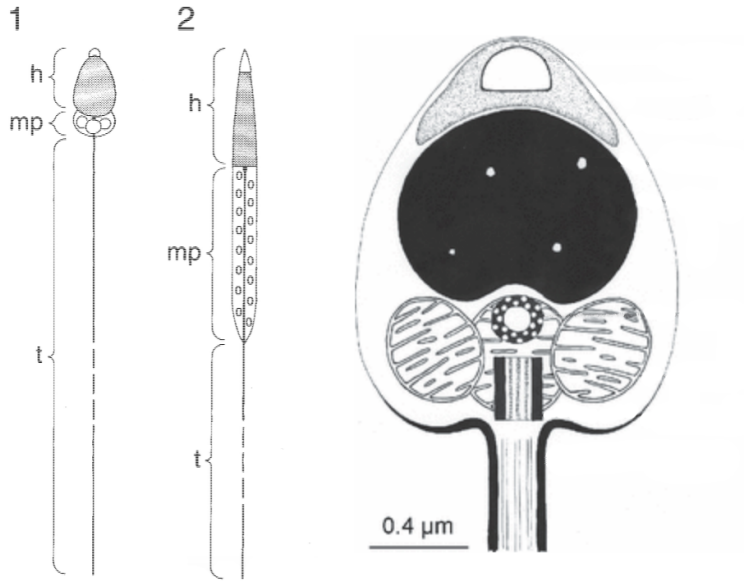


Figure 1.7.: Sperm structure comparison. Left (modified from Lundin and Hendelberg [1998]): (1) "primitive" type of metazoan spermatozoa (2) "modified" type of metazoan spermatozoa - Right (modified from Obst et al. [2011]): spermatozoon of *Xenoturbella* which is an example of the "primitive" type. h - head; mp - middle piece; t - tail.

1.2.2. Molecular analyses

Xenturbella bocki as a mollusc

The first phylogenetic study involving genomic data from *X. bocki* did not match previous groupings based on morphology. *X. bocki* was grouped with Platyhelminthes alongside the Acoelomorpha due to its simple body plan and absence of distinguishing morphological characters. However, the results of the first molecular study rather surprisingly did not to support this and instead linked *X. bocki* with the Bivalvia (phylum Mollusca).

Norén and Jondelius [1997] reconstructed a phylogenetic tree that placed *X. bocki* at a more derived position emerging from within the protostomes. They sequenced the

small-subunit ribosomal RNA gene (18S rRNA) and the protein-coding mitochondrial cytochrome *c* oxidase subunit I gene (COI) from five specimens of *X. bocki*. With a high support (88% jack-knife frequency) they inferred an unresolved clade comprising *X. bocki* and representatives of Mollusca, Annelida, Echiura, Phoronida, Brachiopoda, Entoprocta, Ectoprocta and Nemertea (fig. 1.8). In all jack-knife replicates of the COI data, *X. bocki*'s sister species was the protobranch bivalve mollusc *Ennucula tennis* (fig. 1.9).

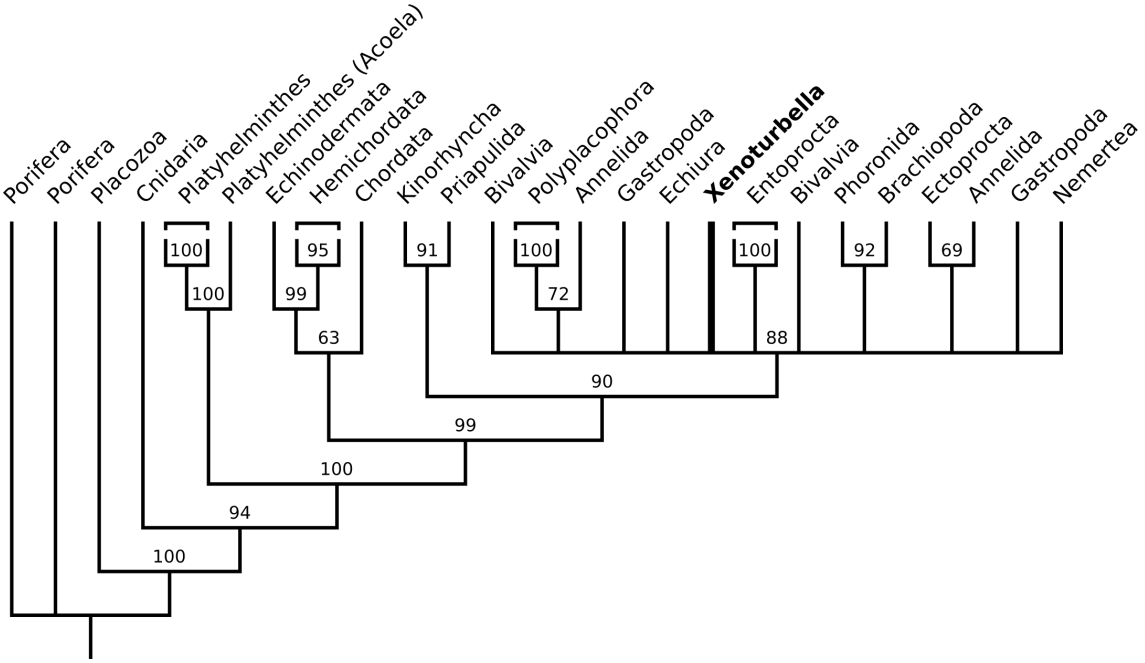


Figure 1.8.: Redrawn from Norén and Jondelius [1997]: first phylogenetic analysis of 18S rRNA data including *X. bocki*. 60% jack-knife replicates consensus tree places *Xenoturbella* within an unresolved clade of Mollusca. Numbers indicate jack-knife frequencies, asterisks mark clades representing multiple species.

There is very little morphological support for a close relation of *X. bocki* and bivalves. Israelsson [1997] described similarities in *X. bocki*'s oogenesis with that of Protobranchia. Israelsson also subsequently claimed that late *X. westbladi* larvae share some characteristics with protobranch bivalves [Israelsson, 1999]. However, he did not observe the typical molluscan arrangement of cells during the cleavage of the embryos (known as "molluscan cross").

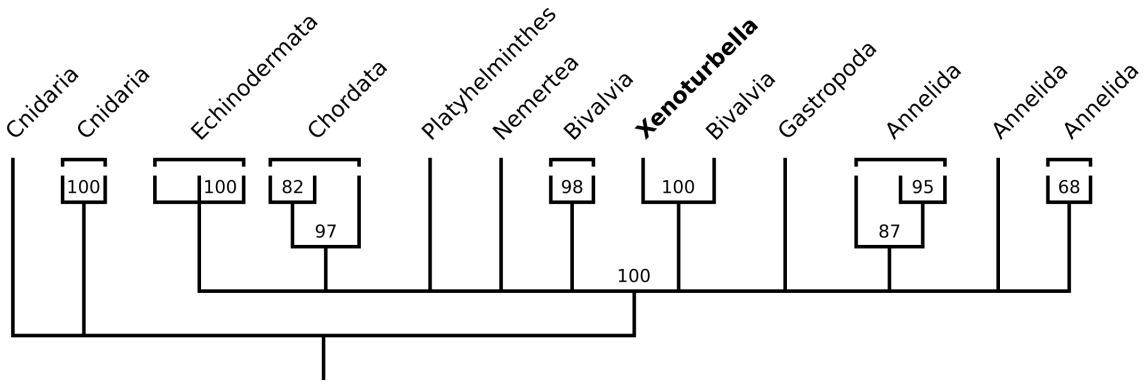


Figure 1.9.: Redrawn from Norén and Jondelius [1997]: first phylogenetic analysis of COI data including *X. bocki*. 60%: jack-knife replicates consensus tree places *Xenoturbella* as sister to Bivalvia. Numbers indicate jack-knife frequencies.

A derived position amongst Bivalvia would raise many questions about *X. bocki*'s evolutionary history. Key bivalve characteristics such as shell, foot, ctenidia, ganglia, digestive tract, heart and circulatory systems are all absent from *X. bocki*. Major loss of characters at this scale would have to be corroborated with genomic evidence.

Bourlat et al. [2003] refuted the bivalve relationship. They stated that only a radical metamorphosis involving the loss of "all bivalve characters" could explain such a connection. In their own experiment (sequencing 18S rDNA, *cox1* and *2*) they did find sequences most similar to those of bivalve molluscs, but also another distinct sequence. By carefully excluding the gut content prior to the DNA extraction they showed that the molluscan signal belonged to ingested prey rather than *Xenoturbella* itself. They supported this claim by showing that the suspicious sequences were almost identical to sequences from molluscs living in the same geographical area [Bourlat et al., 2008].

Xenturbella bocki as a deuterostome

More recent studies found evidence to link *X. bocki* and Ambulacraria (hemichordates and echinoderms). Bourlat et al. [2003] first showed the grouping of *X. bocki* to the Ambulacraria using molecular data. Analysis of the small-subunit (SSU) ribosomal RNA gene showed significantly more support for a position of *Xenoturbella* as sister group of the Ambulacraria than as the sister group of the Bilateria or sister to the Deuterostomia. COI sequences placed *Xenoturbella* as sister group to the hemichordate *Balanoglossus*.

However, as stated by the authors, COI provides less reliability at this level of divergence compared to SSU.

Additional genomic data further supported the inclusion of *X. bocki* among deuterostomes, but its exact relation to other clades was disputed. Bourlat et al. [2006] used another 170 nuclear protein coding genes as well as the complete mitochondrial genome which confirmed their previously inferred phylogeny. However, Perseke et al. [2007] found that the support for *X. bocki*'s position as sister to Ambulacraria hinges on the inclusion of urochordates. After exclusion of the urochordate sequences, *Xenoturbella* was placed at a position as sister group to chordates and ambulacrarians.

Broader genomic analyses strengthened the support for *X. bocki* as sister to Ambulacraria. Dunn et al. [2008] used expressed sequence tags from 77 taxa and 150 genes from a very broad taxonomic range to resolve the relationships within the Metazoa. The results showed high bootstrap support for placing *Xenoturbella* within the Deuterostomia and high posterior probability for placing it next to Ambulacraria (fig. 1.10).

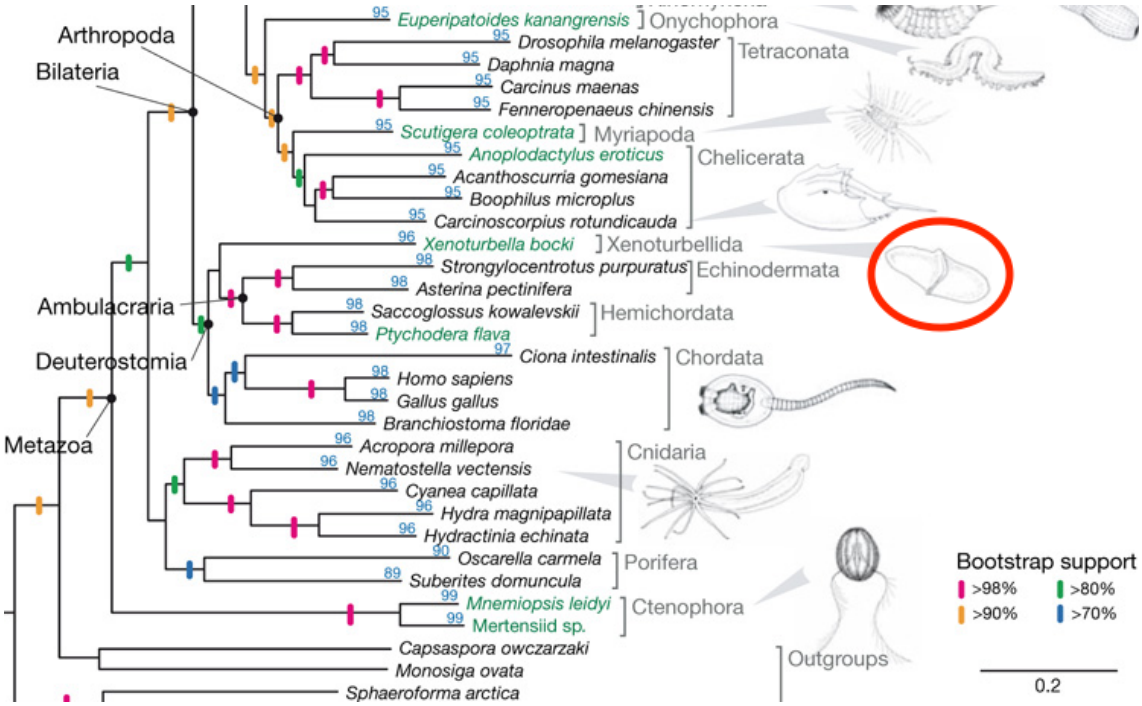


Figure 1.10.: Modified from Dunn et al. [2008]: Tree reconstruction based on 150 genes from broad phylogenetic sampling places *Xenoturbella bocki* (red circle) as sister to Ambulacraria.

X. bocki was also investigated for specific genetic markers to resolve its position within Bilateria. The studies listed above have used a set of genes sampled from wide range of taxa to increase phylogenetic signal and reconstruct the evolutionary history. Other studies have looked at key genetic characters to support, refute or interpret the newly found position of *X. bocki* among deuterostomes.

Hox genes are an important class of genes sharing a common binding motif (the "homeobox") and are involved in the shaping the body plan during embryonic development. In particular, these genes control the patterning of the body plan along the anterior-posterior axis [Lemons and McGinnis, 2006]. A larger complement of Hox genes has been associated with more complex body plans, which could give insight into the interrelationship of distantly related species. Fritzsch et al. [2008] found five *Xenoturbella* sequences containing the homeobox and compared them to Hox genes of other deuterostomes. The number of Hox genes is comparable to those found in Acoela and Nemertodermatida, which is lower than those found in other bilaterians. Assuming their divergence at the base of Bilateria, these findings imply a much simpler hypothetical ancestor for all bilaterian animals than previously thought (fig. 1.5). However, the posterior Hox gene *Xb_HoxP* clusters with the ambulacrarian PG9/10 and chordate PG9 and PG10 sequences, suggesting that *Xenoturbella*'s reduced number of Hox genes could represent an early version of the deuterostome Hox cluster.

Mitochondrial data shows similarities between *X. bocki* and deuterostomes. Hemichordata and Vertebrata share the same arrangement of mitochondrial genes encoding proteins, tRNAs and rRNAs [Castresana et al., 1998] which would represent the state of the mitochondrial genome in the deuterostome ancestor. Perseke et al. [2007] found the same gene order to be conserved in *X. bocki*, but unlike previous studies, their analysis of all 13 protein coding mitochondrial genes did not result in a placement of *Xenoturbella* as sister to the Ambulacraria, but instead a position as sister group to other deuterostomes. Bourlat et al. [2009] also found the ancestral deuterostome mitochondrial arrangement to be present in *Xenoturbella*. Their analysis of inversions and break points showed the closest similarity to hemichordates.They also show how using an improper model of evolution could result in the basal deuterostome position instead of the previously inferred position next to ambulacrarians as it also misplaces urochordates. Testing several phylogenies they concluded that *Xenoturbella*'s position as sister to Ambulacraria could not be rejected using the mitochondrial data.

Mitochondrial gene code places *X. bocki* outside Ambulacraria, not within. Bourlat et al. [2009] found *X. bocki's* mitochondrial genome to be most similar to that of hemichordates. Their analysis of the mitochondrial gene code, however, places *Xenoturbellida* outside the Ambulacraria (hemichordates and echinoderms). Ambulacrarians possess a gene code that deviates from the standard code (e.g. invertebrate mitochondrial code). This genetic code determines which combination(s) of 3 nucleotides (i.e. codon) synthesises which protein. A change from the standard code is seen as a rare evolutionary event and a code shared between closely related species would therefore represent a synapomorphy for this clade [Telford et al., 2000]. *X. bocki* possesses the standard invertebrate mitochondrial genetic code excluding it from Ambulacraria.

1.3. Xenacoelomorpha

The grouping of *Xenoturbella bocki* and Acoelomorpha was already proposed based on morphological similarities, but initial molecular analyses did not group the two clades. *X. bocki* was inferred to be part of Deuterostomia [Bourlat et al., 2003, 2006, Perseke et al., 2007, Dunn et al., 2008] while Acoelomorpha were inferred to be sister to all other bilaterians [Ruiz-Trillo et al., 1999, Jondelius et al., 2002, Ruiz-Trillo et al., 2002].

Hejnol et al. [2009] were the first to infer a monophyletic grouping of Acoelomorpha and *Xenoturbella bocki* based on molecular data using a multi-gene approach. The monophyly of the clade was further supported by Philippe et al. [2011] which proposed the name "Xenacoelomorpha". However, both of these studies fundamentally disagree on the placement of the Xenacoelomorpha within the Metazoa which has been fuelling discussions ever since.

1.3.1. Xenacoelomorpha as sister to Nephrozoa

Morphological support

Most bilaterians possess an excretory system which regulates osmotic pressure and excretes waste products from the organism via nephridia (e.g. vertebrate nephrons that comprise the kidney). The name "Nephrozoa" was first proposed after early molecular phylogenetic analyses showed Acoelomorpha to be sister to all remaining Bi-

lateria [Jondelius et al., 2002]. The name is derived from the presence of nephridia in protostomes and deuterostomes which are absent from acoelomorphs. Despite the absence of dedicated organs, genes involved in ultrafiltration and nephrocyte structures have been identified in Xenacoelomorpha [Robertson, 2017].

Haszprunar [2016] favours a position as sister to Nephrozoa and lists several characters that would contradict a derived placement amongst deuterostomes (or ambulacrarians in particular). (1) Differences in musculature (epithelial in most deuterostomes, fibrous in Xenacoelomorpha) makes a derived position unlikely. (2) All ambulacrarian larval types feature an apical organ which is absent from xenacoelomorphs (although they do not have a larva). (3) Metanephridia are absent from xenacoelomorphs. (4) Deuterostome larvae feature coeloms in early development, but no trace of coelomic cavities can be found in Xenacoelomorpha.

Molecular support

The first molecular analysis to unite the Xenacoelomorpha [Hejnol et al., 2009] used a 1487 genes matrix containing 94 taxa including *Xenoturbella bocki* and 6 acoelomorphs. The goal of the study was to assess the relationship of Bilateria focusing on Acoelomorpha in particular. They inferred a position for Xenacoelomorpha at the base of all Bilateria.

Srivastava et al. [2014] investigated the regeneration in the acoel *Hofstenia miamia* and used its transcriptome to infer the phylogenetic relationship to other bilaterians. They argue that *H. miamia* is a good acoel candidate for phylogenetic analyses as it shows a slower molecular rate compared to other acoels [Jondelius et al., 2011]. Together with publicly sourced data from *Nemertoderma westbladi* and *Isodiametra pulchra* they found acoels to be positioned as sister to all other bilaterian animals. When they included data from *X. bocki* it disrupted their inferred topology and lowered support for basal branching acoels, but also other clades.

Cannon et al. [2016] used several models to show a robust inference of Xenacoelomorpha as sister to Nephrozoa. They used 212 orthologous groups, determined the best fitting model for tree construction and supported the findings through additional Bayesian analyses. The authors suggest that previous placement of Xenacoelomorpha within deuterostomes [Philippe et al., 2011] was caused by insufficient data and reliance on ribosomal proteins.

1.3.2. Xenacoelomorpha as a sister group to Ambulacraria

In contrast to the basal bilaterian position, Philippe et al. [2011] inferred the Xenacoelomorpha as sister to the Ambulacraria (hemichordates and echinoderms) using a more sophisticated model to reduce the effects of systematic errors. This position is in agreement with earlier investigations that included *X. bocki*, but no acoelomorph species [Bourlat et al., 2006, Dunn et al., 2008] They showed that using less fit models to infer the phylogenetic tree disrupts its topology which would explain a shift of Xenacoelomorpha towards a more basal position. They observed a high evolutionary rate of xenacoelomorphs compared to other bilaterians. Inferring phylogenetic relationships when dealing with large differences in molecular change between species can lead to problems in tree reconstruction, such as Long Branch Attraction (LBA) [Felsenstein, 1978]. LBA causes organisms with high evolutionary rate to be inferred closer to each other than their true phylogeny, e.g. inferring fast evolving organisms to be closer to outgroups than they actually are.

Morphological support

Morphological support for a position amongst Ambulacraria has already been described in *X. bocki*' s first description by Sixten Bock, stating that the "body walls of *Xenoturbella* and those of Enteropneusta are so similar that one could think *Xenoturbella* could belong to the balanoglossids" (cited in Westblad [1949]). Reisinger described *X. bocki*'s statocyst to be very much like those of Synaptidae (a family of sea cucumbers) and its epidermis to be nearly identical with those of enteropneusts [Reisinger, 1960]. Franzén and Afzelius [1987] noticed a pattern of cilia and rootlets similar to *Xenoturbella* in the pharyngeal cilia of enteropneust *Glossobalanus minutus* (fig. 1.6).

Criticism about the morphological similarities between *X. bocki* and Ambulacraria were voiced several times: Ehlers and Sopott-Ehlers [1997] argued that the statocyst structures are not homologous, whereas epidermis similarities are merely superficial (in agreement with Pedersen and Pedersen [1988]). Shared characters in cilia and rootlet structures [Franzén and Afzelius, 1987] were argued against by Lundin based on the fact that the domed distal cap is different [Lundin and Hendelberg, 1998] (fig. 1.6). Analysis of myocytes also proved inconclusive: while their shape and junctions are similar to

those found in some species of Hemichordata, the inner lamina to which the myocytes are anchored in those species is missing in *Xenoturbella* [Ehlers and Sopott-Ehlers, 1997].

Molecular support

Philippe et al. [2011] analysed 197 genes and found xenacoelomorphs to group with Ambulacraria. They also used mitochondrial proteins, which resulted in a position at the base of the deuterostomes and an unresolved relationship between chordates, ambulacrarians and xenacoelomorphs. In addition, they sequenced a microRNA in both acoels and *Xenoturbella*, which had only otherwise been found in deuterostomes (miR-103/107/2013) and a microRNA in *Xenoturbella*, which had previously only been found in ambulacrarians (miR-2012). Further support came from a deuterostome specific sperm protein (RSB666) they identified in xenacoelomorphs.

The gene GNE is exclusively encoded in deuterostomes (except for urochordates where it was secondarily lost). de Mendoza and Ruiz-Trillo [2011] showed that GNE is encoded in deuterostomes, acoelomorphs and *Xenoturbella*, but not in protostomes or non-bilaterians. Despite this, they hesitated to use it as a phylogenetic character. They argued that a Xenaceolomorpha clade basal to all bilaterians and a gene loss in the lineage leading to the protostome ancestor could be a viable explanation.

1.4. Phylogenetic implications

1.4.1. Xenacoelomorpha as sister to remaining Bilateria

A position of the Xenacoelomorpha as the sister group to all other bilaterian animals affects our understanding of the characteristics of both the ancestor to all Bilateria and the common ancestor of Protostomia and Deuterostomia (PDA). Features shared between all bilaterians, but not outside are novelties gained after the split from the common ancestor with the Cnidaria. Features present in Protostomia and Deuterostomia, but not Xenacoelomorpha are gained after their respective divergence from the common ancestor.

Several morphological features can be straightforwardly inferred as present in the ancestor of all Bilateria. The existence of a posterior-anterior and a dorso-ventral axis in

both xenacoelomorphs and all other bilaterians implies that this development in body shape already existed in the common ancestor. Another common character is the existence of the mesoderm germ layer which appears during early development of the embryo. The mesodermal cells later differentiate to form muscles.

Morphological characters found in protostomes and deuterostomes but not xenacoelomorphs were gained after divergence of the PDA from the bilaterian common ancestor (unless these features have been lost in xenacoelomorphs). Xenacoelomorphs do not possess a dedicated organ or system for excretion (i.e. nephridia), a fact that led to naming the remaining Bilateria "Nephrozoa". Other organs, such as circulatory or respiratory systems absent in xenacoelomorphs, may also have existed in the nephrozoan ancestor. Furthermore the ancestor of nephrozoans most likely possessed a through gut different from the xenaceolomorphs' blind gut. While there are more derived bilaterians that share a lack of a through gut (e.g. Platyhelminthes) this would be explained by a secondary loss rather than a convergent evolution of the through gut in most other bilaterian lineages. While concentration of nerve fibres (CNS) seems to have evolved in both acoels and nephrozoans [Achatz and Martinez, 2012], ganglia or a true brain are also an apparently novel feature of the Nephrozoa.

On a molecular level we can also trace changes and elaboration of genetic features after the xenacoelomorphs' split from the remaining Bilateria. The Hox/ParaHox cluster in Acoelomorpha is smaller than other bilaterians [Jiménez-Guri et al., 2006]. The expansion then would have occurred in the lineage leading to the nephrozoan ancestor. Similarly, there was an expansion of the miRNA complement after the split from Xenacoelomorpha explaining the lack of bilaterian miRNA families in acoels [Sempere et al., 2007].

Based on these inferences it would be possible to make assumptions about the ancestor of all bilaterians. The urbilaterian (ancestor of all Bilateria) was most likely a directly developing (i.e. without larval stage) acoelomate worm featuring a blind gut with a single opening [Haszprunar, 2016]. It would have lacked ultrafiltration cells, true eyes and ganglia.

Given this assumption, it would be possible to identify key elements (genes, gene regulatory networks) that play into the development of specific tissues and organs that evolved after the split from the rest of the Bilateria.

1.4.2. Xenacoelomorpha as a derived clade within Bilateria

A position of Xenacoelomorpha within the Bilateria as sister to Ambulacraria implies a more complex urbilaterian. The common ancestor would still feature above mentioned characters such as bilateral symmetry and three germ layers. Additionally it would have most likely possessed coelomic cavities, organs, respiratory and circulatory systems which are prevalent in both Protostomia and Deuterostomia (compare with PDA in fig. 1.5.

Assumptions about the deuterostome ancestor would most likely not be affected by a derived position of the Xenacoelomorpha. The urdeuterostome most likely possessed a foregut with gill slits and an excretory system as this is common to both Chordata and Ambulacraria [Nielsen, 1995]. There was most likely no centralised brain as the nervous system differs between the notochords in chordates, the non-homologous stomochord of the hemichordates and the weakly centralised nervous system in ambulacrarians.

A xenacoelomorph position as sister to Ambulacraria implies morphological simplification and the loss of deuterostome characters in the Xenacoelomorpha. The absence of protonephridia, gill slits, through-gut and other bilaterian or deuterostome characters has to be explained by secondary loss.

Comparisons with other known cases of simplification (e.g. tunicates, platyhelminths) could reveal mechanisms of character loss. Pathways and gene-gene interactions which are linked to specific morphological traits need to be analysed and examined in species that arguably lack these characters. Furthermore, it may answer questions about the importance and potential compensation for organ systems absent from these phyla.

1.5. Thesis objectives

1.5.1. Establish a method to scrutinise orthology inference to improve clade specific sets of orthologues and apply it to Bilateria, Protostomia and Deuterostomia

What impact does gene absence have? The interpretation about importance and effect of genes needs context. We must establish what genes are comparable and how we can compare them. We need information about how genes are related to each other and where these related genes are present. Taxonomically restricted genes could represent

unique adaptations. A common presence of certain genes implies an evolutionary pressure to preserve their sequence and function. The phylogeny of organisms that show no presence of a gene could imply an ancestral state of their genome or the loss in the lineage leading to the extant organisms.

The simple morphology of Xenacoelomorpha represents an interesting case study to compare it to other bilaterians also on a genetic level. Regarding the xenacoelomorphs' hypothesised positions either at the root of the Bilateria or within the Deuterostomia, it is most interesting to see differences to bilaterian and deuterostome specific characters.

To establish clade specific characters we need to carefully combine and validate information from a large range of species. In this project I will describe my approach to infer clade specific genes of high quality. I will address common pitfalls of these kinds of investigations that we have identified and propose solutions to minimise their impact.

The identification of shared genes relies on the accurate inference of homologous gene relations. When investigating conserved gene functions we need to identify which and how genes are related.

Introduction to homology, orthology and paralogy

Shared characters can have two origins: a shared common ancestry or convergent evolution. Shared common ancestry implies the existence of the shared characters in the most recent common ancestor between the species compared (also called plesiomorphy). Morphologically shared characters are often used as descriptions of the clade, e.g. the Chordata possess a notochord and Mollusca have soft bodies. Convergent evolution, in contrast, represents adaptations in two taxa that did not exist in the most recent common ancestor of the two taxa. This can be observed in species that have independently adapted to the same or similar lifestyles such as the development of wings in Chiroptera (bats and flying foxes), Aves (birds) and Pterygota (winged insects).

Characters that are similar due to common descent (were already existent in the most recent common ancestor) are called homologous ("same relation"). The term has been adapted to also describe genetic characters that can be found in several species that conclude a shared ancestry. Akin to using shared morphological characters, genetic characters are used to infer and describe the hypothetical most recent common ancestor that features these characters.

Genes can have more than a simple ancestor-descendant relationship which lead to the distinction of several types of homologous genes [Koonin, 2005] (fig. 1.11). a) Orthology describes related genes that diverged after a speciation event. It is generally assumed that orthologous genes retain their original function based on the observation of a lower evolutionary rate than other forms of homology. b) Paralogy describes related genes that diverged after a gene duplication event. The existence of a second copy could facilitate one copy to acquire new functionality (neofunctionalisation) or both copies to divide the original function (subfunctionalisation). This is supported by the higher evolutionary rate observed in paralogous genes. c) Xenology describes related genes diverging after a horizontal gene transfer. Unlike orthologous and paralogous genes the most recent common ancestor did not feature the gene in question. d) Ohnology is a special case of paralogy, where gene copies occurred through a whole genome duplication rather than a singular gene duplication event. e) Homoeology in a polyploid organism describes the relationship of homologous genes that were brought together by a hybridisation event of previously diverging lineages.

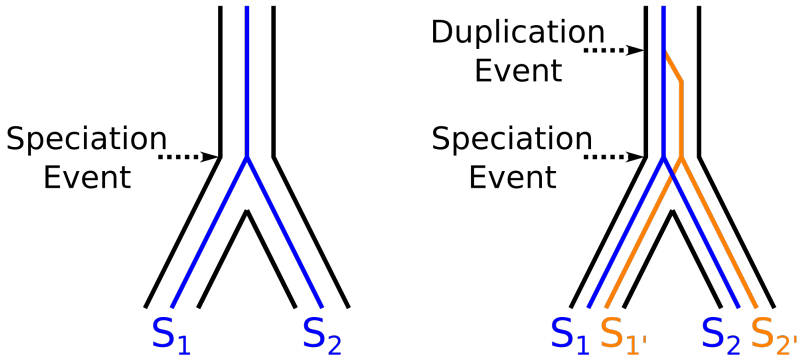


Figure 1.11.: Different types of homology show how related genes diverged from a common ancestor. Orthologous genes (S_1, S_2) diverged after a speciation event, paralogous genes (S_1, $S_{1'}$) diverged after a gene duplication event.

The most important types of homologous genes in phylogenetic studies are orthologues as they directly correlate with speciation events. This makes them useful for the inference of phylogenetic trees to resolve species relationships and estimate divergence times (using molecular clocks as a way to measure changes over time). They are also important to

investigate presence or absence of specific gene functions. Shared gene presence can unite clades as a defining character. Ancestral absence can be used as a parsimonious measure to resolve early branches within a clade while secondary loss can determine great shifts within a subclade through the loss of function.

The identification of orthologues can be non-trivial and needs careful consideration. Orthology is typically inferred by sequence similarity. For protein coding genes the amino acid sequence is used to avoid overestimation of differences caused by synonymous changes, i.e. nucleotide changes that do not affect the protein sequence (through different codons coding for the same amino acid). Protein sequences also allow to see changes in their domain architecture. Shared protein domains are an indicator for a shared gene function.

Sequence similarity of orthologous genes requires the use of thresholds which can be problematic in distantly related or fast evolving lineages. In order to find orthologues, sequences are typically aligned and compared by how well they match each other. Through the accumulation of mutations over time these sequences can diverge past a certain point where similarity can be detected. The same problem can occur if one or both sequences display a high evolutionary rate that causes a higher dissimilarity. Using thresholds to determine when genes are still considered similar is a nontrivial issue especially when trying to include species of both varying evolutionary rates and relationship distance.

Compounding the challenges with orthology identification, paralogy can be hard to distinguish from orthology in the presence of gene loss. Paralogous genes are related through a gene duplication event. The time point of the gene duplication (in relation to speciation events) can be obscured if one or more copies between species were affected by gene loss. The existence of two copies in one but not the other species could infer a gene duplication event in the species possessing both copies or a duplication in the shared ancestor followed by a loss of one copy within one lineage. Even more problematic is the case of hidden paralogy, where two species contain one copy each, but these stem from a gene duplication event which preceded the loss of one copy each in both lineages.

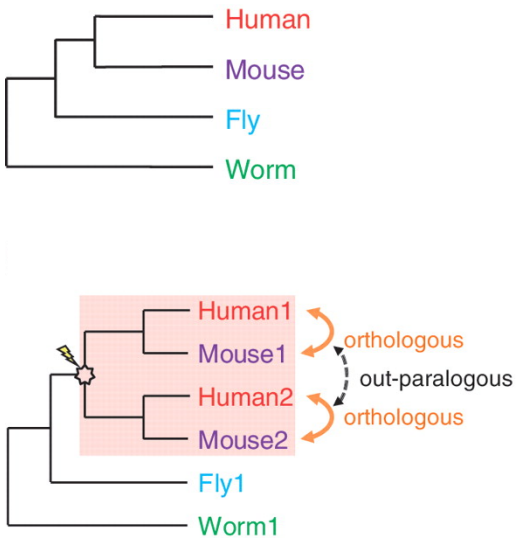
Orthology inference methods

A plethora of tools and methods to accurately identify orthologous genes has been developed (see Kristensen et al. [2011] for a non-exhaustive review which lists 17 published

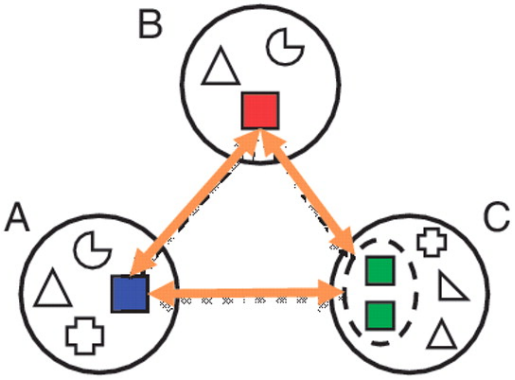
methods). The existence of such a high number of methods which is continuously growing showcases the complexity of the underlying problem. Altenhoff and Dessimoz [2009] compared several orthology inference projects and methods. They tested how well the inferred orthologous genes reconstruct the species tree and how similar these genes are in their putative function. They conclude that choice of method depends on the trade-off between sensitivity and specificity. Different projects also provide different coverage (included taxa) of precomputed orthologous genes, which can be beneficial depending on a study's scope. Lastly, they emphasise that a simple bidirectional best hit (BBH) approach shows a good overall performance compared to more sophisticated methods. BBH approaches are usually ad-hoc implementations to find a pair of genes that are most similar to each other than any other gene in the respective organisms. This usually involves the application of similarity scoring methods, e.g. BLAST [Altschul et al., 1990], to find putative orthologous genes for several species and then investigating the established set of genes.

Most orthology inference methods either use tree based or heuristic best match methods to identify orthologous genes. Tree based methods infer a gene tree which represents how homologous genes are related. The methods then try to reconcile the gene tree with a tree representing the relationship of the species involved (fig. 1.12a). The species tree can be either given (if known) or inferred based on the information gathered by all inferred gene trees (e.g. by identifying a consensus tree supported by a majority of gene tree relations). Under the assumption of a correct species tree these methods can lead to accurate estimation of gene duplication/loss events and relationships. The most problematic case is when the species tree is based on problematic data which is then in turn used to infer the gene relations it was based on. Heuristic best match methods avoid this by inferring orthology based purely on gene-gene similarity without regards for an underlying species tree. This makes these methods less biased, but more computationally demanding. Instead of reconciling gene trees with a species tree, heuristic best match methods generate graphs connecting genes based on their relatedness (approximated through their pairwise similarity). These graphs are then clustered and divided to find groups of orthologous genes (fig. 1.12b).

I was involved in the development of OMA standalone (see Appendix C.3), a standalone application of the OMA project (Appendix C.1, Altenhoff et al. [2016]). OMA and OMA standalone have been used in several projects since its publication. I was part



(a) Inference of orthology and paralogy using a species tree (top) and a gene tree (bottom). Homologous genes that follow the gene tree (e.g. Human1, Mouse1, Fly1 and Worm1) are orthologous as the diverged after speciation events. Two copies of the ancestral gene in both humans and mice imply a gene duplication event (lightning bolt) in the ancestor of human and mouse.



(b) Inference of orthology and paralogy using a graph based approach. The graph visualises the relation between individual genes (shapes) in different species (circles). A clustering method identified the genes that are most similar to each other (squares connected by arrows). These genes form a cluster of orthologues, called orthogroup. The orthogroup contains a gene duplication that only occurred in the lineage of species C (dashed circle).

Figure 1.12.: Differences between orthology inference approaches (modified from [Kristensen et al., 2011]).

of a project that used OMA to analyse the genetic turnover in bats (Appendix C.2). OMA is a graph-based approach to infer orthologous relationships. OMA applies extra steps to ensure that orthology inference is as free from false positives as possible. One such check is the establishment of verified pairs, i.e. pairs of most similar sequences between two species. These pairs are tested for hidden paralogy, i.e. one of the sequences is actually paralogous, but differential gene loss caused the paralogous sequence to be the only copy left [Altenhoff et al., 2016]. This increased specificity comes at the cost of a decrease in sensitivity [Altenhoff et al., 2016]. OMA also provides the option to generate Hierarchical Orthologous Groups (HOGs). HOGs are clusters of related genes at a given taxonomic level. HOGs use a species tree (given or inferred by OMA) to represent not only orthologous and paralogous relations, but also the gene events (gain, loss and duplication) that lead to them.

Using several orthology methods can lead to increased scope, but also to conflicting results. I will demonstrate how we can integrate different results to capture orthologous relations inferred by one method that may have been overlooked by other methods. I also deal with conflicting results and how it is possible to include them if there is at least partial agreement between different methods.

An accurate identification of clade specificity hinges on the taxa included. One problem we have identified is taxon sampling bias which can negatively impact claims to clade specificity. Taxon sampling bias occurs when species are not included in an investigation that could later be used to disprove the inferred clade specificity. This phenomenon is born out of the limitations of computational analyses. With the amount of data currently available, including all sequenced species is not feasible.

Results of orthology inference derived from a relatively limited set of taxa should be tested against much more comprehensive databases to reduce potential taxon sampling bias. As a very simple check to validate my results for bilaterian, protostome and deuterostome specific orthologues I searched for potential homologues outside their respective clades. This approach will show me if my orthologues truly are novel genes and only existent within my clades of interest or if taxon sampling bias obscured the relationship to species outside.

Finding potential homologues for clade specific orthologues does not necessarily invalidate the inferred results. Sequence similarity alone is not a predictor for orthology, i.e. conserved gene function. If all potentially homologous sequences outside the specified

clade are diverging in a paralogous manner clade specificity can be retained. I will explain how I use information from gene trees generated from the orthologous gene sets and the found similar sequences to identify paralogy.

Which genes are specific to Bilateria, Protostomia and Deuterostomia, and what functions do they have? After inferring and validating the specificity of orthologous gene sets I will make predictions about their potential functions and how they may have impacted the emergence of their respective clades.

1.5.2. Establish a method to use known microRNA families and small RNA sequence data to search for microRNA evidence in Xenacoelomorpha

Complex structures and their progenitors arise during the early development of animals. The lack of complex structures in Xenacoelomorpha could be a result of a partial or complete absence of the underlying gene network. Here I focus on on class of genes, called microRNAs, that have been been associated with the emergence of organ systems and specialised tissues [Heimberg et al., 2008].

Introduction to microRNAs

MicroRNAs (miRNAs) are an abundant class of short non-coding RNA molecules of about 19-22 nucleotides that are involved in RNA silencing and post-transcriptional regulation of gene expression. They not only appear in animals, but plants as well. In plants each miRNA has one target mRNA while miRNAs in animals typically have several targets. Their main mode of gene regulation involves binding to the target mRNA which creates a double-strand pairing. This results in a decrease in gene product through a) blocking the ribosome translating the mRNA into a protein and b) faster degradation of the mRNA (fig. 1.13)

In plants, miRNAs match their target mRNA (near) perfectly. This results in a one-to-one relationship between a miRNA and the gene it affects. The miRNA hybridises with its matching mRNA which leads to the cleavage of the mRNA transcript and prevents translation into a protein [Jones-Rhoades et al., 2006].

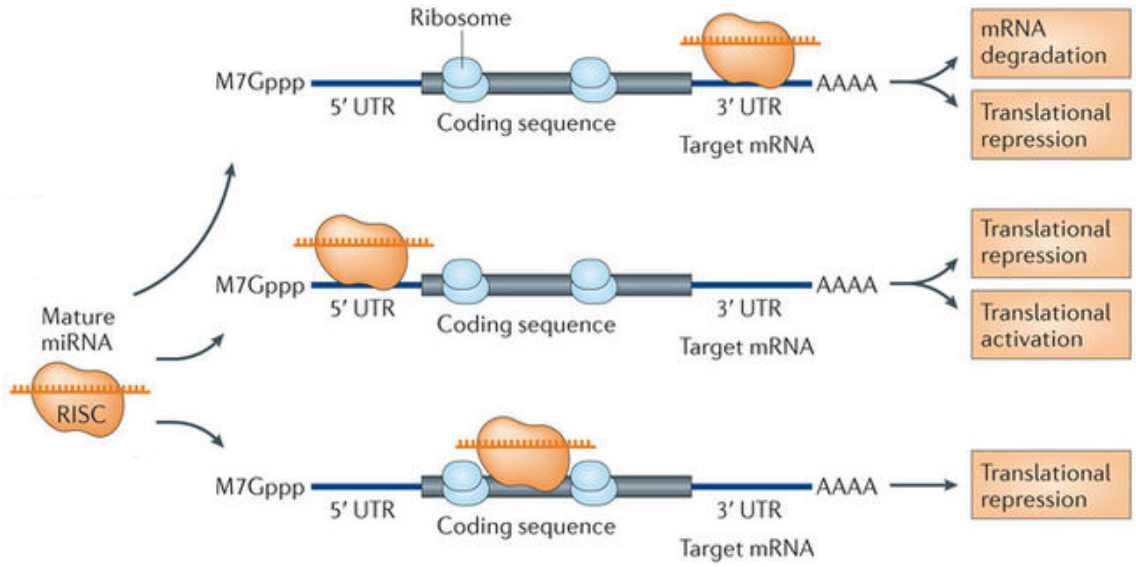


Figure 1.13.: Mechanism of microRNA suppressing gene expression, modified from Ling et al. [2013]. The mature miRNA gets incorporated in the RNA-induced silencing complex (RISC). The RISC binds to the miRNA's target and blocks full or partial translation of the protein.

In animals miRNAs can each target more than one mRNA. This leads to a more complex relationship between a miRNA and all its potential targets. The main deciding factor about a miRNA's potential effects is its "seed" region. This region is located at the 5' end of the mature miRNA sequence (usually starting at nucleotide 2) and is 6-8 nucleotides long. The seed region determines which targets a miRNA will bind to, but its short length allows it to bind to several mRNAs. The potential for several targets allows for a much greater combination of interactions where one miRNA can affect regulation of several mRNA targets and, vice versa, one mRNA transcript may be regulated by several microRNAs.

In animals miRNA's seed region is perfectly conserved. The seed region determines which genes are affected by a given miRNA. A single nucleotide change could result in a drastic change of binding potential to target mRNAs. Based on this microRNA families are distinguished by their members' seed regions. Seed regions that match perfectly between miRNAs of two animals have the potential to affect the same target genes (if present).

Core components of the miRNA pathway are conserved between plants and animals, but their respective miRNA repertoires have emerged independently [Shabalina and Koonin, 2008]. As we are investigating the xenacoelomorphs' relationship to other animals I will focus exclusively on the description of miRNAs in animals in the rest of this thesis.

MicroRNA biogenesis

MicroRNAs are subject to a complex process that requires a specific structural composition at certain stages during its biogenesis (fig. 1.14).

The mature (i.e. acting) microRNA is part of the primary microRNA (pri-miRNA) which is transcribed from the genome. Rather than being transcribed directly, mature miRNAs are part of this larger sequence. Pri-miRNAs can contain a cluster comprising several mature miRNAs (e.g. *mir-23∼27∼24-2* cluster [Lee et al., 2002]). The pri-miRNA folds to form a hairpin structure for each mature miRNA to allow for subsequent cleavage.

Hairpin structures of the pri-miRNA are cut and released for further processing. Each hairpin structure comprises a double-stranded stem region (where the sequence hybridises to itself) and the unpaired hairpin loop region. The stem region is recognised by nuclear proteins Drosha (in vertebrates) or Pasha (invertebrates) which then cut ahead of it. This freed hairpin structure is termed the precursor microRNA (pre-miRNA) which is then transported from the nucleus into the cytosol.

In the cytoplasm the pre-miRNA hairpin structures are processed into the mature microRNA. Mature miRNA sequences are about 22 nucleotides in length and embedded in the stem region of the pre-miRNA hairpin. The RNAse III enzyme Dicer cleaves the hairpin structure to remove the loop sequence and part of the stem region resulting in a double-stranded miRNA duplex. The non-acting strand of the duplex gets degraded while the mature miRNA gets incorporated into the RNA-induced silencing complex (RISC) where it will interact with its target mRNA.

MicroRNAs as phylogenetic markers

Expression of miRNAs differs between cell types and tissues and impacts developmental and other biological processes [Bartel, 2004, Wienholds et al., 2005]. As such, miRNAs

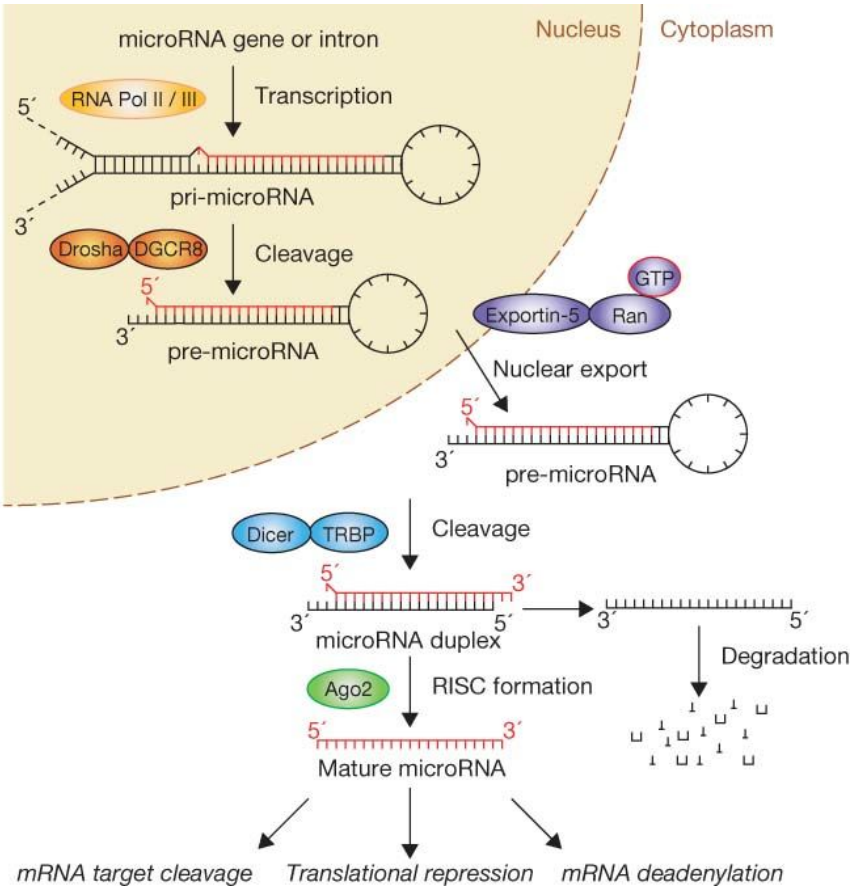


Figure 1.14.: Biogenesis of microRNA, from Winter et al. [2009]. Pri-microRNA gets transcribed from the genome. The enzyme Drosha cleaves the pre-microRNA hairpin structure from the pri-microRNA. The pre-miRNA is exported from the nucleus and subsequently cleaved by the enzyme Dicer. The microRNA duplex dissociates and the acting strand is incorporated into the RNA-induced silencing complex (RISC) while the inactive strand is degraded.

are an important marker to look into when exploring the presence or absence of morphological characters.

lin-4 was the first miRNA discovered, but only the widespread abundance of *let-7* led to the classification of miRNAs as distinct regulatory elements. *lin-4* was first sequenced in the nematode *C. elegans* [Lee et al., 1993, Wightman et al., 1993]. *let-7* was later also found in other bilaterians but not in Cnidaria, Ctenophora or Porifera [Pasquinelli et al., 2000, 2003]. Pasquinelli et al. [2003] hypothesised a connection between *let-7* expression

and terminal differentiation of structures such as organs, tissues and specific cell types. This would indicate *let-7* being associated with the emergence of new morphological characters within the Bilateria.

MiRNAs are highly conserved across different species. Several studies have shown miRNAs to be preserved between distantly related species [Berezikov et al., 2005, Sempere et al., 2006, Zhang et al., 2006]. Wheeler et al. [2009] estimated that miRNAs evolve more than twice as slowly as 18S rDNA which has been used in phylogenetic studies due to its high conservation amongst metazoans.

MiRNAs are continuously added, but rarely lost [Hertel et al., 2006, Sempere et al., 2006, Prochnik et al., 2007]. Research in metazoans has shown that the number of miRNAs correlates with the taxonomic hierarchy of animal relationships [Sempere et al., 2006]. The observed purifying selection of miRNAs leads to a rapid fixation, high conservation and little loss. The strong selection is attributed to the role of miRNAs in gene regulation. A major change or even loss of a single miRNA could affect several target genes causing potentially detrimental effects.

MiRNAs have been linked to morphological complexity. Sempere et al. [2006] highlights the correlation between the estimated number of cell types and the estimated number of miRNAs in *H. sapiens* and *D. melanogaster* lineages over time. Heimberg et al. [2008] shows a correlation between the rate of miRNA family acquisition and the vertebrate morphological complexity.

Presence of miRNA families has been used to resolve difficult cases of phylogenetic relationships. The high rate of conservation avoids issues caused by gene loss or high molecular rates. Sempere et al. [2007] used information about bilaterian and platyhelminth miRNAs and their absence in acoels to infer a position of Acoelomorpha as sister to all other Bilateria. Phylogenetic studies of cyclostomia (lampreys and hagfish) [Heimberg et al., 2010] and Mandibulata (e.g. crustaceans, insects, millipedes) [Rota-Stabelli et al., 2011] used miRNA data to support their monophyletic grouping. Campbell et al. [2011] also used miRNAs to resolve the relationship between Tardigrada, Onychophora and Arthropoda.

More recent analyses contradict the notion of an ever growing miRNA repertoire. Fromm et al. [2013] have analysed platyhelminths and found many miRNA families missing that are otherwise conserved across the Bilateria. Platyhelminths are notorious for their change in phylogenetic position throughout history (see above). Their now

established position amongst the Lophotrochozoa implies a reduction from a more complex ancestor. The loss of ancestral miRNAs could be causal to the simpler body plan compared to sister taxa. Thomson et al. [2014] reanalysed previous miRNA studies and found issues with naïve parsimonious approaches. Their investigation showed that loss of miRNA families is more widespread than previously believed (up to 54% of families affected in Bilateria). They explain the shortcomings of simple parsimony methods given these new insights and propose the use of more sophisticated Bayesian statistical methods to estimate phylogenies.

MicroRNA detection in Xenacoelomorpha

The first studies have shown an absence of many bilaterian miRNA families in acoels. Sempere et al. [2006] identified 20 miRNAs conserved within Bilateria. They tested 16 families, but were only able to sequence 6 in the acoel *Childia* sp. They repeated the analysis for *Symsagittifera roscoffensis* and reached the same conclusion [Sempere et al., 2007].

The first studies have shown an absence of many bilaterian miRNA families in acoels. Sempere et al. [2006] inferred 17 miRNAs to be conserved between humans and fruit flies (*D. melanogaster*), with one more miRNA added based on a previous publication [Aboobaker et al., 2005]. As these miRNAs are shared between a protostome and a deuterostome, they are inferred to have existed in the ancestor to all bilaterians. An extended search in more bilaterian species revealed 2 more miRNAs which are absent from humans and fruit flies, respectively, increasing the total amount of miRNA families to 20. They sequenced only two of these miRNAs (*miR-10* and *miR-100*) in cnidarians and none in sponges. They tested 16 miRNA families on the acoel *Childia* sp., but only found 6 miRNAs to be present. They regarded the absence of many of the conserved bilaterian miRNAs as support for a sister relationship between xenacoelomorphs and the remaining bilaterians. Sempere et al. [2007] re-examined the phylogenetic distribution of bilaterian miRNAs. They were also unable to sequence more than the previously found bilaterian miRNAs in the newly included acoel *Symsagittifera roscoffensis*. This re-enforced the notion of a widespread absence of miRNAs shared between protostomes and deuterostomes in acoel species.

Follow-up investigations of *Xenoturbella bocki* and acoel *Hofstenia miamia* showed a higher prevalence of bilaterian miRNA families compared to previous studies. In *H. miamia* Philippe et al. [2011] were able to sequence 10 of the miRNA families missing from *Childia* sp. and *S. roscoffensis*. They identified an additional 8 bilaterian miRNA families in *X. bocki*. Together with the inferred slower evolutionary rate of Xenoturbellida compared to its sister clade, their findings indicate a slight loss of miRNAs in the ancestor of Acoelomorpha.

Deuterostome specific were found in both *X. bocki* and *H. miamia*. Further support for the xenacoelomorphs' position amongst deuterostomes stems from the identification of miRNA family *mir-103/107/2013* which is specific to Deuterostomia (also Philippe et al. [2011]).

Data for miRNA detection in Xenacoelomorpha is currently insufficient. Most miRNA detection prediction programs (see reviews in Bentwich et al. [2005], Bortolomeazzi et al. [2017]) require not only the sequences from potential miRNA candidates, but also supplementary information such as miRNA information about closely related species (e.g. *miRseeker* [Lai et al., 2003] or Berezikov et al. [2005]) or require the method to be trained on comparable data (e.g. *PalGrade* [Bentwich, 2005] or *HHMMiR* [Kadri et al., 2009]). The contentious position of Xenacoelomorpha makes it difficult to choose ideal organisms to compare to. The lack of good quality sequencing data exacerbates the problem.

MiRNA data acquisition in Xenacoelomorpha is problematic and requires careful investigation for potential candidates. If the observed higher absence of miRNAs in acoelomorphs is a result of loss rather than a reflection of an ancestral absence, they may not accurately represent the state of miRNAs in the xenacoelomorph ancestor. Unfortunately, *X. bocki* is notoriously hard to sample and sequence (in communication with other members of the group).

In this project I will establish how we use sensitive RNA extraction methods and draft genome information to identify miRNA candidates. Together with Peter Sarkies, our collaborator at Imperial College London, we wanted to explore the possibility to use small RNA extracts and information about pri-, pre- and mature miRNA structures to establish a pipeline for miRNA detection. The pipeline will be applied to small RNA and draft genome data of *X. bocki*.

To increase the scope of our investigations I will also showcase how to predict miRNA candidates from genome information alone. This allows us to also find candidates in acoelomorph draft genomes for which we currently do not have small RNA sequences and compare it to the findings from *X. bocki*.

I also validate my methods against generated and real data. The pipeline is tested against generated data to test the rate at which miRNA candidates could be erroneously identified. As a positive control I also apply my pipeline to high quality genomes to estimate the rate at which I can identify established findings in these organisms.

2. Establishing high confidence core orthologous gene sets

2.1. Motivation

Our main focus is the investigation of similarities and differences between Xenacoelomorpha and the remaining bilaterians. The simple morphology of Xenacoelomorpha has to be investigated in light of genetic innovation amongst other bilaterian clades. As such xenacoelomorphs represent a pivotal taxon to investigate the evolution of bilaterally symmetric animals. The phylogenetic position of Xenacoelomorpha is still in debate (see Introduction and Telford and Copley [2016] for a description of the problems in placing this taxon). In either case we have to explore the absence of other bilaterian characters. The observed complexity of most other bilaterians such as internal organs, blood vessels and body patterning could be caused by genes or gene interactions which are missing from the Xenacoelomorpha.

The comparison of Xenacoelomorpha to the rest of the Bilateria requires the establishment of robust sets of genes which are specific to their respective clades. Currently, two hypotheses have been proposed for the placement of Xenacoelomorpha within the animal tree of life: a) sister to all remaining Bilateria and b) sister to Ambulacraria (hemichordates and echinoderms) within Bilateria. A sister relationship to all remaining Bilateria implies that molecular changes had accumulated between the ancestor of Xenacoelomorpha and Bilateria and the ancestor of Protostomia and Deuterostomia. Genes conserved between protostomes and deuterostomes may be absent from xenacoelomorphs. We are interested in finding genes specific to Bilateria which are absent from Xenacoelomorpha. If xenacoelomorphs are part of the deuterostomes, we are also interested in deuterostome specific genes. The placement within Deuterostomia implies a loss of many deuterostome characters in the lineage leading to the xenacoelomorph an-

cestor. The analysis of deuterostome specific genes could reveal which genes are absent from xenacoelomorph species.

In this chapter I will present my approach to infer and curate a set of genes specific to a given clade of animals. Clade specific gene sets have been published in the past [Simakov et al., 2015, Krämer-Eis et al., 2016], but I have found issues in the methodologies used. The goal of my project is to address these issues and propose solutions to create robust sets of clade specific genes by decreasing the found biases in previous work. These sets of genes are meant to be compared to either a species or clade of interest.

This chapter will explain how I use the relationship of genes to establish clade specificity and validate these findings. I will explain what kinds of gene events lead to different kinds of gene relationships and how this affects what can be considered specific. I will showcase how I identified issues with previous findings and use this information to create validity checks for my own inference results. I will establish a pipeline that follows the aforementioned principles to create robust gene sets. In the chapter afterwards, I will apply my pipeline to Bilateria, Protostomia and Deuterostomia, creating a set of clade specific genes for each group of animals.

2.1.1. Gene events throughout evolution

From their emergence to their presence in extant species genes can have a varied history. Specific gene events and evolutionary restrictions lead to different evolutionary paths. It is important to untangle the complex relationships of genes to gain insight into the emergence or change of genes over time and what their presence or absence implies for their respective lineages.

The first step to understand the relation of genes is to find out which genes are homologous, i.e. share a common ancestor. This allows us to separate genes that share an evolutionary history from those that do not. If genes share a common ancestor, we can trace their lineage throughout the tree of life and analyse significant events. These gene events can result in different types of homologous relations. The most important distinction is between homologous genes that are orthologous and those that are paralogous. Orthologous genes have diverged from a common ancestor after speciation events, i.e. the genetic history of these orthologues is congruent with the phylogenetic history of

the species involved. Paralogous genes have diverged after a gene duplication event, i.e. two or more copies of the ancestral gene started diverging from one another.

We can use sequence similarity between related genes to build gene trees, i.e. a reconstruction of the genetic history of homologous genes. Together with a phylogenetic tree, i.e. a reconstruction of speciation events, we can map the occurrence of genes within clades of organisms and infer how the gene complement has changed over time. Depending on their prevalence we can distinguish several classes of gene events that affect the different lineages:

Gene gain (also: emergence or novel gene event) describes the emergence of a gene in the stem lineage of a clade (fig. 2.1 G). The inference of gene gain events follows from the absence of the respective gene from all outgroup species. Novel genes may occur *de novo* from a previously untranscribed genomic region through the acquisition of a promotor or other regulatory sequences (via point mutations, insertions or deletions of nucleotides). Through selective pressure these genes became fixed and were propagated. Another reason for inferring novel gene events is the lack of similarity to related genes, i.e. the ancestral sequence diverged rapidly from a common ancestor to the point where we are unable to detect the relationship to other genes. Due to the lack of homologues found in other species novel genes are also called "orphan genes" which could represent important lineage specific adaptations [Tautz and Domazet-Lošo, 2011].

Gene duplication can be inferred from the existence of several paralogues within a clade. If this amount differs from a clade's respective sister clade, one or more rounds of gene duplication must have occurred within the lineage leading to the clade's ancestor (fig. 2.1 D). Duplications can be facilitated through transposable elements or even the duplication of the whole genome. The duplication of a gene is assumed to cause an increased divergence of the two paralogous copies [Koonin, 2005]. One reason could be the potential for neo- or subfunctionalisation: neofunctionalisation allows one copy to adopt a new function, providing additional benefits without interfering with the original copy, while subfunctionalisation results from both copies dividing the original gene's function between them. Duplications are the source of expansion of related genes leading to gene families such as the *Hox* gene family or MAPK.

Gene loss follows from the absence of a gene in a species or clade of interest while the gene has been inferred to exist at an ancestral level (fig. 2.1 L).

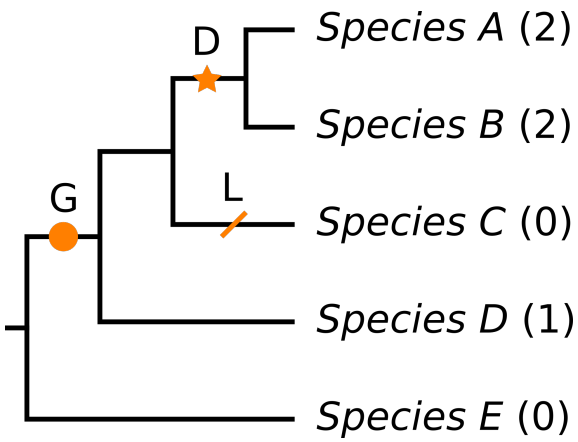


Figure 2.1.: Phylogenetic species tree showcasing different gene events. The number of homologous genes present in each species is stated in parentheses.: **G** - gene gain after split from common ancestor with *Species E*; **D** - gene duplication leading to two paralogous copies in *Species A* and *B*; **L** - gene loss in *Species C*.

We can use these inferred gene events to establish which genes are unique for our clades of interest. Genes gained at the most recent common ancestor could represent clade specific adaptations. We want to investigate what differentiates Bilateria and Deuterostomia from their respective outgroup species and relate that information to what we can or cannot find in Xenacoelomorpha.

2.1.2. Previous work

In this section I will present previous studies to establish clade specific genes in Bilateria and Deuterostomia. In the following section I will state our concerns with the methods that have been used so far. There I will present explicit cases that refute or weaken the presented claims and argue how I will use this information within my own approach to avoid or diminish these shortcomings.

Simakov et al. [2015] present new draft genomes for two hemichordates and use this data to compare with other bilaterians. Their main focus for the gene comparison was the inference of deuterostome specific novelties, i.e. new genetic traits that evolved in the lineage leading from the ancestor of Bilateria to the ancestor of Deuterostomia (chordates, hemichordates and echinoderms). For their newly assembled genomes they used transcriptomic data to annotate genes that could then be compared to annotated

genes in other bilaterians. The authors used BLAST, a method to calculate how similar sequences are to each other, to generate scores for all pairwise comparisons between sequences. They used these scores and a custom clustering approach (described in their previous publication [Putnam et al., 2007]) to group genes into clusters of orthologous genes. This approach used sequence similarity to find ingroup sequences across different species that are most similar to each other. From these scores they constructed clusters of orthologous genes based on the most similar gene pairs. Outgroup sequences were then used to merge separated clusters if the outgroup sequences were amongst the most similar sequences compared to more than one ingroup cluster.

They present deuterostome novelties divided into 4 different types depending on their relation with homologous sequences found in non-deuterostome species: **Type I** novelties of are gene novelties of unknown origin, i.e. they did not find any homologous sequences outside the deuterostomes. **Type II** novelties introduce new protein domains, that they did not find in non-deuterostome homologues. **Type III** novelties introduce new protein domain architectures, i.e. the order of protein domains differs from homologous non-deuterostome sequences. **Type IV** novelties were inferred to have an accelerated rate of evolution relative to non-deuterostome homologues, i.e. they inferred significantly more molecular change in the lineage leading to the deuterostome sequences than to non-deuterostome sequences. For each potential novelty they required at least 2 species to be present from each deuterostome superphylum, i.e. 2 or more chordates and 2 or more ambulacrarians (Hemichordata+Echinodermata).

We are most interested in type I novelties as these represent genes exclusive to the Deuterostomia. Simakov et al. [2015] report more than 30 of these deuterostome specific orthologous groups. The hypothesised position of Xenacoelomorpha as sister to the Ambulacraria puts the focus on the differences of xenacoelomorphs and the deuterostome ancestor. The inferred groups are very useful to search for genes that are shared with Xenacoelomorpha as well as finding genes which are absent from Xenacoelomorpha.

Krämer-Eis et al. [2016] searched for genes unique to Bilateria to find a potentially common genetic basis for the traits that are shared across bilaterian species. They report 85 clusters of orthologous genes that are associated with the development of body plan, nervous system and muscles, as well as cell-cell communication. To establish these clusters, they also used BLAST to calculate how similar the analysed sequences are to each other. They excluded all bilaterian sequences for which they were able to find

potential orthologous sequences in non-bilaterians. The orthology was inferred by a reciprocal best hit approach, i.e. two sequences within two species being more similar to each other than any other sequence within these species. If one of the species involved was not part of the Bilateria, but contained such a sequence, the bilaterian sequence was excluded from further analysis. This step guaranteed the bilaterian specificity of the genes kept. The resulting set of bilaterian specific sequences was clustered using the orthology inference methods OrthoMCL [Li et al., 2003] and InParanoid [O'Brien et al., 2005]. These clusters represent orthogroups, groups of genes that are orthologous to each other and therefore likely share similar functions. Each cluster was analysed using gene ontology (GO) enrichment analysis to identify their putative functions.

Paps and Holland [2018] analysed the gene content of a large variety of animals to reconstruct the ancestral metazoan genome. Their focus was mainly on inferring gene numbers and the associated gene gain and loss events. They first searched for homologous genes (using the BLAST similarity scores and the Markov clustering approach). Then they mapped the found clusters using a custom script onto the topology indicating when gene groups emerged and when they were subsequently lost. This information allowed them to estimate the number of genes in the various common ancestors across the Metazoa. The functional analysis focused specifically on the novel groups and those lost in the ancestor of all animals. The analysis was done using the fruit fly representatives and their GO annotation.

2.1.3. Problems identified in previous work

BLAST

The inference of homology requires a way to identify how closely related sequences are. The relatedness of gene sequences is typically expressed via overall similarity of the sequence sites (nucleotides for DNA/RNA, amino acids for proteins). BLAST [Altschul et al., 1990] is a commonly used tool to calculate the similarity of two sequences. The relatedness is typically expressed via the expect value (*e*-value) which represents the chance of finding a similar sequence within a given database. The lower the *e*-value the more significant the similarity between the query sequence and the matched sequence, and the lower the chances of identifying these two sequences to be similar by chance.

Identifying a suitable *e*-value for finding related sequences is non-trivial. Homologous sequences from closely related species are expected to be more similar than homologues of more distantly related species. Additionally, sequences can have different divergence rates. A higher mutation rate (e.g. due to environmental pressure) will cause a sequence to evolve faster which results in lower similarity scores. The sequence could then appear more distantly related than its true evolutionary relationship (e.g. Long Branch Attraction).

Simakov et al. [2015] used a custom pyramidal method to cluster sequences which involved the computation of reciprocal best BLAST scores between sequences. Their threshold for clustering was 10^{-4} [Putnam et al., 2007]. However, they used a "moderate (10^{-20})" *e*-value to check if their inferred "deuterostome novel gene families" are similar to any non-deuterostome sequences. We deem this threshold to be too stringent for relating sequences spanning ∼650 million years of evolution. I did a BLAST search of sequences which were reported as "novel without non-deuterostome homologues" (Group G1 in supplementary material) and found several hits in non-deuterostome species (most notably in *Drosophila*) albeit at higher *e*-values of less or equal to 10^{-5}. In one particular case (group 174191) I was able to identify potential homologous sequences across the tree of life (fig. 2.2).

Both Krämer-Eis et al. [2016] and Paps and Holland [2018] used BLAST with an *e*-value of 10^{-5} to assess sequence similarity. While there is no consensus on the best suited *e*-value, I have noticed that many studies involving a wide range of species (e.g. spanning several kingdoms or even domains) have used 10^{-5} arguably as a compromise to identify related sequences of distantly related species at the expense of misidentifying non-related sequences in closely related species.

In my approach, we decided to use the default *e*-value of BLAST (10) to vastly increase sensitivity in finding related sequences. While this will undoubtedly increase the risk of erroneously finding false positives, I apply strict validity checks on my results in order to minimise their impact.

Taxon sampling

In order to classify which genes are represented by which clade we need a sufficient amount of species represented in our analyses. If a clade is underrepresented it can lead

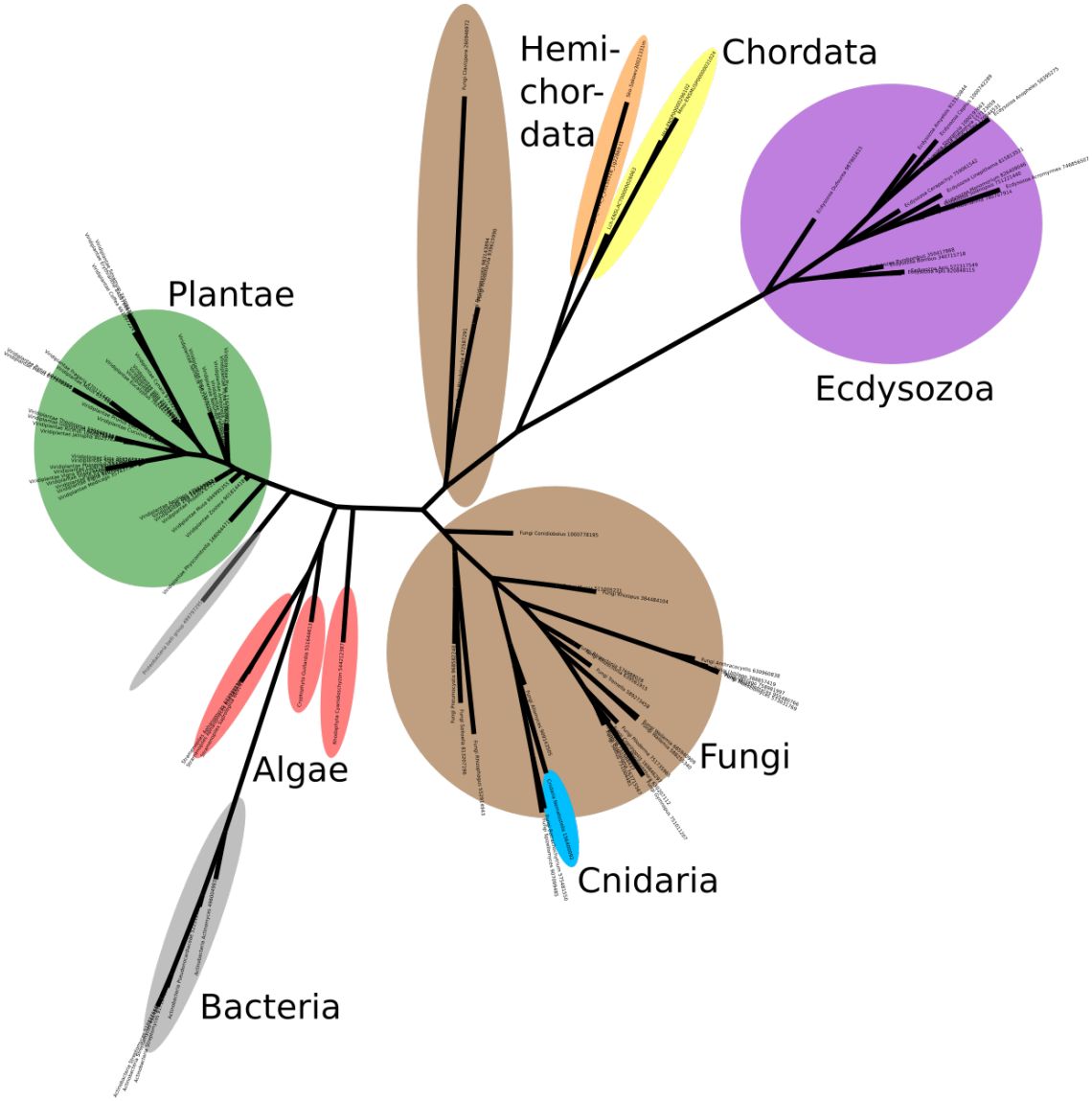


Figure 2.2.: Gene tree reconstructed from potential homologues for Simakov et al. [2015] group 174191. Group 174191 was classified as "gain type I, with no BLASTP hit outside of deuterostomes" and contained only sequences from Chordata (yellow) and Hemichordata (orange). I found many potential homologues outside the deuterostomes also using BLASTP. Annotations describe the found sequences as "Ribosomal protein L33". The tree reconstruction of these sequences shows a widespread presence in all domains of life and good phylogenetic separation.

to an overestimation of absence (or potentially loss) of genes. Lineage specific losses are propagated to losses or gene absence for the whole clade. The overestimation of absence can lead to the misidentification of gene origins and the inference of clade specific losses.

Due to computational restrictions it is not possible to include all currently available genomic data. As mentioned above, the inference of sequence relations requires the comparison of all sequences against each other. As computing time increases exponentially with regards to the number of sequences compared, it is not possible to include all available species that have currently been partially or fully sequenced. A careful curation of the used dataset is important to find a suitable trade-off between a fair representation of analysed clades and the resources available to complete the analysis in a timely manner.

Any result inferred from a limited set of species should be scrutinised to minimise the effects of taxon sampling bias. Simakov et al. [2015] used BLAST to query the sequences they deemed deuterostome novelties "of unknown origin" in order to find potential non-deuterostome homologues. I assume that the database in question is NCBI's non-redundant protein sequences (nr) as this is the BLAST default and there is no mention to the contrary in the supplementary notes. In their analysis Krämer-Eis et al. [2016] included only a single species for many of the non-bilaterian clades (two in Cnidaria). A lineage specific loss in any of these given species would be interpreted as an absence for the whole clade. They did not apply any check to the results of their analysis to remedy this. Paps and Holland [2018] only validated the group of Novel Core HG (homologous representatives found in all or all but one species comprising a given clade). They used the fruit fly representatives and performed a BLAST search against the NCBI GenBank database.

A manual check of the data presented by Simakov et al. [2015] resulted in contradictory findings. Our interest is focused on *Xenoturbella* and its potential affiliation to the deuterostomes. As a routine check I used the groups of deuterostome novelties and searched for potential homologous sequences in a large set of protostome sequences curated by Max Telford (dataset accessible via https://doi.org/10.5281/zenodo.2650166, access embargoed until 24/04/2020). I was able to identify potential non-deuterostome homologues for several of the reported groups.

Our concerns about the validity of the reported clade specificity led me to include a validity check against a comprehensive database to filter for false positives that were

caused by a potentially insufficient amount of taxon sampling. After identifying potential clade specific orthologous groups my approach searches the NCBI Reference Sequence Database (RefSeq) to find potential homologues outside the specified clade. I then use these sequences to either confirm or reject the previously established clade specificity.

Orthology inference method

The identification of orthologous relationships requires the careful investigation of the similarity between sequences across species. There is a plethora of different orthology inference methods available [Kristensen et al., 2011] showcasing the difficulties in finding an optimal way to accurately infer orthology.

Simakov et al. [2015] used a custom pyramidal approach that relates genes from the leaves towards the root of their phylogenetic tree. If the genes of ingroup species are more similar to each other than to those of outgroup species the sets of the involved genes get merged into one cluster. This is repeated for every node of the given topology towards the root. In their supplementary material they note that their method accurately segragates different Wnt subfamilies in vertebrates which get merged at more ancient nodes (Bilateria, Metazoa). This algorithm was used previously [Putnam et al., 2007, Simakov et al., 2013], but has not been compared to other widely used orthology inference methods. We are concerned about the performance and the universal applicability of this approach as we do not have any benchmarks to compare.

Krämer-Eis et al. [2016] used OrthoMCL [Li et al., 2003] to infer orthology. OrthoMCL uses similarity scores to generate a graph (nodes describing sequences, edges describing their similarity) and clusters similar sequences using the Markov Clustering (MCL) algorithm. The resulting set of clusters represents groups of orthologous genes. The species represented in each cluster specify the phylogenetic node at which this orthologous group of genes originated. OrthoMCL is widely used, but was shown to be outperformed by OrthoFinder in both identification of true positives and avoiding false negatives [Emms and Kelly, 2015]. OrthoFinder works similar to OrthoMCL, but adds ways to reduce sequence length biases found in OrthoMCL's approach.

Paps and Holland [2018] did not infer orthologues, instead opting to use the MCL algorithm directly on their BLAST scores to generate clusters. These clusters are deemed homology groups (HGs) and were mapped to a given tree topology using a custom script.

They count at each taxonomic level the numbers for total amount of HGs, novel HGs (no representatives outside this level), core novel HGs (novel HGs present in all or all but one species at this level) and lost HGs (present outside this level, but lost in ancestor). They were able to recover traditional gene families/classes/superfamilies (e.g. *Iroquois* gene family or Wnt ligands) and use this to support the findings using this novel approach. They present no comparison to other established methods.

Choosing one inference method over another or even creating a novel approach can introduce biases that might alter the final results (example shown in fig. 2.3). There are efforts to create services to compare your own orthology inference method to several established methods, e.g. the orthology benchmarking service (http://orthology.benchmarkservice.org, Altenhoff et al. [2016]) which uses a set of well established orthogroups to compare against.

In order to reduce a single method's bias and showcase differences in orthology detection, my approach uses 3 different orthology inference methods: i) OrthoMCL [Li et al., 2003] which is widely used (currently at 2943 citations) and provides results comparable to Krämer-Eis et al. [2016], ii) OrthoFinder [Emms and Kelly, 2015] for its improvements over OrthoMCL, and iii) OrthoInspector [Linard et al., 2011] for its higher sensitivity in inferring orthologous relations. Despite my involvement in its development, we decided against the inclusion of OMA as another orthology inference method. OMA is known for its high specificity, but low sensitivity compared to other methods [Altenhoff et al., 2016]. The lack of sensitivity would cause problems when trying to find orthogroups comparable to those of the other methods. I will explain how I integrate the partially differing results of the three included methods to reach an agreed set of orthogroups.

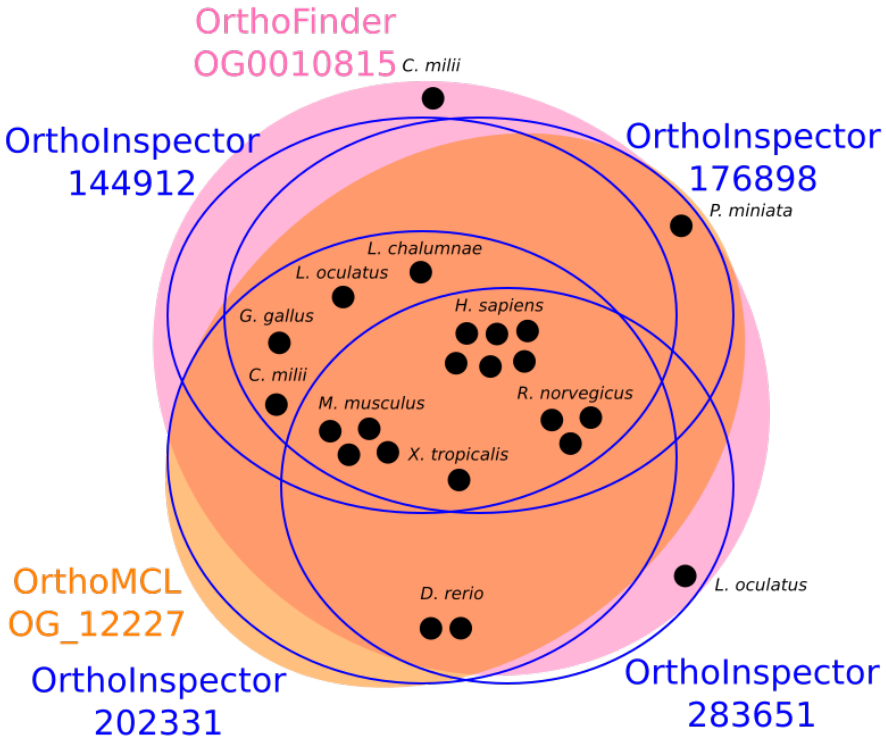


Figure 2.3.: Different orthology inference methods result in different orthogroups. Black circles represent related genes of the MTMR1 gene family. OrthoFinder (pink) includes two genes from *C. milii* and *L. oculatus*, which are excluded from OrthoMCL's grouping (orange). OrthoInspector (blue) identifies four overlapping sets of orthologous genes depending on which gene is used to start the orthology inference. None of the OrthoInspector groups contain the *C. milii* sequence which is only included by OrthoFinder.

2.2. Material

We selected 36 species across the Metazoa and 3 choanozoans to include in this study. Most metazoans were chosen based on their status as "model organisms" which usually guarantees a higher quality of genome data available. We supplemented this initial dataset with a few additional species to improve representation for each clade of the Bilateria, e.g. adding *P. miniata* to match the number of echinoderm genomes to those

of hemichordates and increase the total number of ambulacrarians (see fig. 2.4).

Most data were retrieved from public databases such as NCBI, Ensembl, Ensembl-Genomes and Uniprot. This was complemented by genomes such as *Patiria miniata*, *Hypsibius dujardini*, *Adineta vaga* and other species from more specialised sources to even out representation over the tree (sources for all genomes are listed in Appendix A). Unfortunately, as of writing this thesis, there is a lack of model organisms and therefore good quality genomes outside the Bilateria. The impact of this undersampling will need reassessment in the future.

OrthoMCL (see below) includes a filtering step to get rid of sequences too short to be useful for its orthology inference. I applied this filtering using the default parameters to our dataset which left 1,195,160 protein sequences to be processed by my pipeline.

2.3. Methods

I used DIAMOND (version 0.8.18.80, default parameters, Buchfink et al. 2015) to compute the similarity scores between all protein sequences against each other.

For the orthology inference I used OrthoMCL (version 2.0.9, default parameters, Li et al. 2003) , OrthoFinder (version 0.7.1, default parameters, Emms and Kelly 2015) and OrthoInspector (version 2.21, default parameters, Linard et al. 2011) all using the same filtered protein sequences and DIAMOND results.

I used my OrthoMerge pipeline (described below) to merge the results of all orthology inference methods. I then applied two rounds of validity measurements to check the resulting orthogroups for their clade specificity.

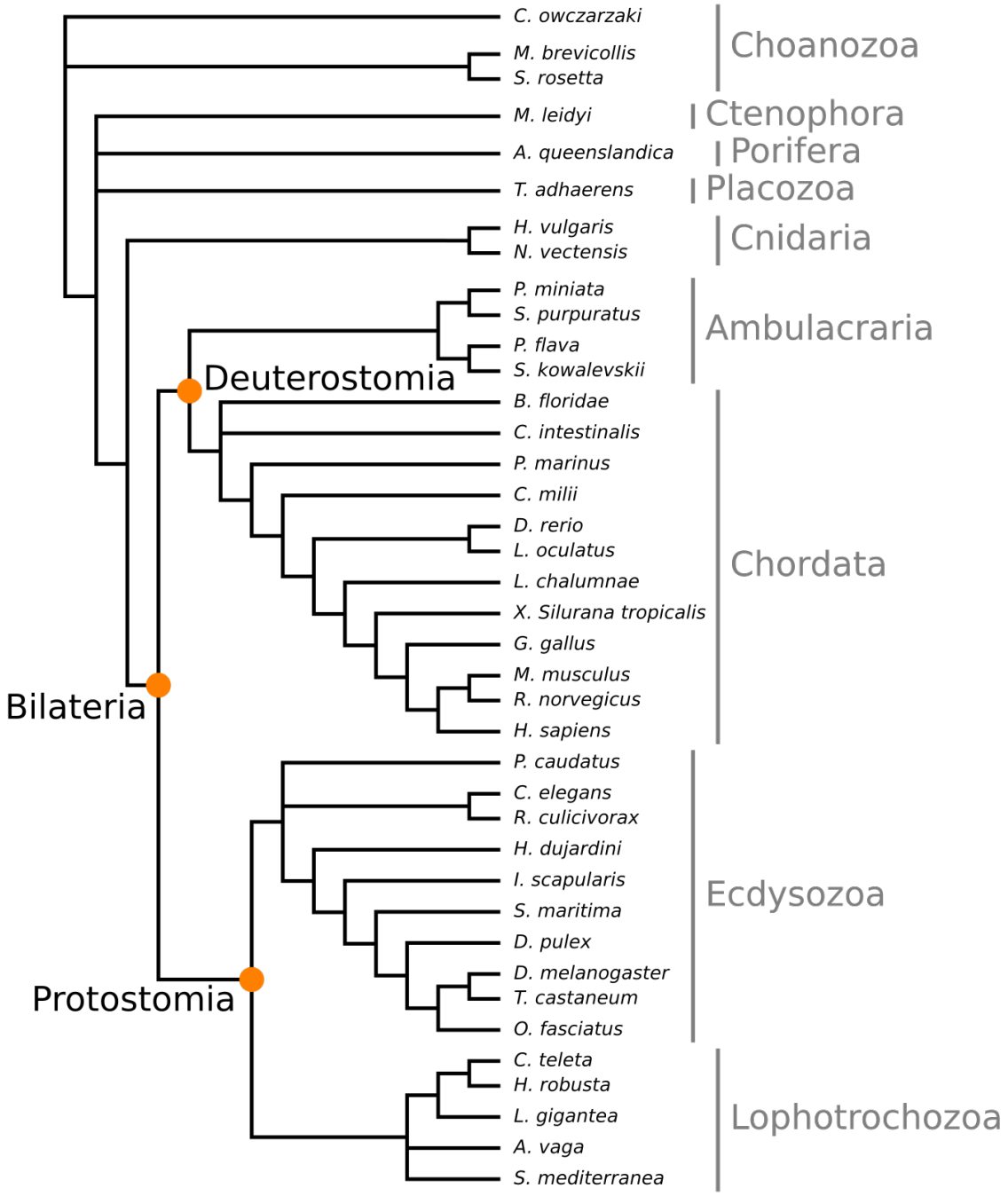


Figure 2.4.: Species and their clades represented in my approach to identify orthologous groups specific to Bilateria, Protostomia and Deuterostomia.

2.4. The OrthoMerge pipeline

2.4.1. Merging different orthology predictions into an agreed secondary set of orthogroups

The use of different orthology prediction methods results in different sets of orthogroups (groups of orthologous genes). These results comprise my **primary set of orthologous genes**. Due to inherent biases of each method only some of these orthogroups will be identical between any two methods. For many groups that are not identical, we can find that they differ to greater or lesser extents, e.g. due to one method being more sensitive in identifying orthologues or more stringent when grouping. A decreasing threshold for grouping orthologous sequences leads to larger clusters. A method with a greater threshold may split these large clusters into smaller clusters, if the connection between the small clusters is below the threshold. The first step of my pipeline is an approach to integrate these differences to find a useful compromise to avoid the exclusion of too many orthogroups.

A first naïve approach is to only use clusters that have been identified as being exactly the same by all methods (see fig.2.5 A). This would be a good initial set, as the clusters' reliability has been shown through identification by several methods. However, this would ultimately be biased towards the most stringent method that produces smaller clusters. The resulting clusters may exclude true orthologous sequences that are not similar enough to be part of the inferred orthogroup, thereby resulting in a large number of false negatives. Both OrthoFinder and OrthoInspector attempt to curb this issue by being more sensitive when compared to older methods such as OrthoMCL (described in their respective publications).

To increase the scope and to reduce the number of excluded groups, I added a condition that would allow a group to be included, if there was partial agreement between the methods. I define partial agreement as orthologous groups that are properly split in other methods, i.e. a group identified by a more sensitive method may be split into proper subsets (and perhaps a number of unassigned sequences) in a different more stringent method (see fig. 2.5 B). My reasoning is that the split of the group occurred due to a lack of sensitivity in the corresponding method. The less sensitive/more stringent method in question would have "failed" (when compared to a more sensitive method) to

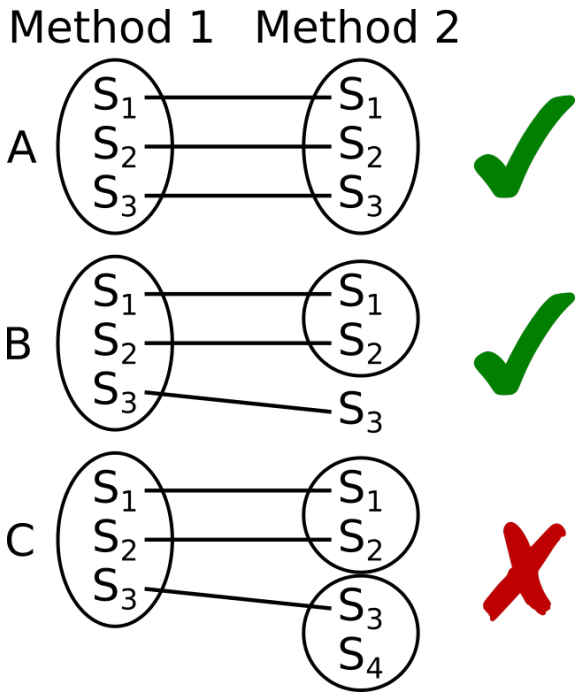


Figure 2.5.: My merging approach uses three cases of (dis-)agreement between orthology inference methods to consolidate their results: **A**: Orthogroups have been identified identically and will be kept - **B**: Orthogroups of one method are split into one or more proper subsets and unassigned sequences, i.e. sequences that were not assigned to any orthogroup. The largest group will be kept. - **C**: Method disagreement leads to overlapping groups that I reject from further analysis. – S_n - Sequence of species n, circles - identified orthologous groups.

identify the connection between the subgroups or the link to more derived and therefore unassigned sequences.

Any remaining groups I consider to be in disagreement. The differences in these clusters could not be solved by uniting smaller groups into bigger groups as this would include sequences that are not part of the bigger cluster (see fig. 2.5 C).

To find partial agreement between methods I compared the results in a bidirectional fashion where a cluster inferred by one method had to agree according to the aforementioned conditions, but also vice versa. This means, that if two groups differ between methods and only overlap partially, that agreement can only be reached if all sequences that do not belong to the intersection of these groups are singletons, i.e. do not belong

to another group of orthologous sequences. I compared between all pairs of methods and only groups that were agreed upon in each pairwise comparison were included in my **secondary set of orthologous genes** (implementation: `OrthoMerge.py`, see Appendix B.1).

2.4.2. Validation and filtering of secondary orthogroups

The secondary set of orthologous groups represents all putatively clade specific orthogroups that were identified by all orthology prediction methods given our selected dataset and the sequence comparison method. We used a wide array of species across and outside Bilateria to reduce the potential for missing orthologues that could show an apparently clade specific orthogroup to be in fact non-clade specific.

I wanted my final sets of clade specific orthologues to be as conservative as possible. I used several inference methods to increase the validity and robustness of my initial set of orthogroups. In the motivation for this project I showed that taxon sampling bias can be an issue for inference interpretation, e.g. sequences from taxa excluded from the analysis can disprove a claim to clade specificity. For each of my found orthogroups I search the NCBI RefSeq database [O'Leary et al., 2016] to search for potential homologous sequences that could disprove clade specificity. I use these sequences to validate my findings by applying two measures: a) I reconstruct a gene tree from an orthogroup and the found RefSeq sequences and check if the orthogroup forms a monophyletic clade. A non-monophyly implies that the orthogroup would have included these sequences, if they had been included in the initial inference. b) I check the found RefSeq sequences for reciprocal best bi-directional hits among the orthogroup sequences, which is a sign for orthology which could disprove clade specificity of my orthogroups.

I subsampled the RefSeq database (downloaded on 09/09/2016) according to our clades of interest. The NCBI RefSeq database contains sequences from more than 62,000 organsims. For each clade, I created a set of genes of outgroup genes. I created a set of $>$61 million non-bilaterian sequences to search for potential homologous sequences to my bilaterian specific orthogroups. Likewise, I created sets of $>$68 million non-protostome sequences and $>$63 million non-deuterostome sequences to find sequences that could potentially refute the claim to clade specificity for the respective clades.

I queried each group of clade specific orthologous genes against the complementary set of the group's respective clade (e.g. a deuterostome specific group was queried against all non-deuterostome sequences). I recorded each hit I found (using BLAST with default parameters, e.g. *e*-value = 10) as a potential homologue to the orthogroup's member sequences. I restricted the results to the best 100 hits found for all member sequences of an orthogroup.

In some cases I was not able to identify any potential non-clade homologues for specific orthogroups (see next chapter). If the analysis was correct, these cases would represent *de novo* gene families that emerged in the lineage leading to the root of their respective clade. These genes are of particular interest as they represent unique lineage specific adaptations.

Monopholy of orthogroup sequences to validate clade specificity

Gene trees from orthologous sequences must reflect species phylogeny. By definition, orthologues are related sequences that diverged after a speciation event, meaning that a reconstructed gene tree using orthologous sequences is congruent to the underlying species tree. Clade specific orthologous sequences must therefore also form a monophyletic clade within the gene tree. Any homologous sequences from species outside the clade in question must be placed outside the corresponding clade in the gene tree. If this requirement is not met, the inferred orthogroup is invalid.

I devised a method to validate an orthogroup's monophyly automatically (implementation: `OG_monophyly_test.py`, see Appendix B.1). From the orthogroup and its potential homologues I built multiple sequence alignments (clustal-omega, version, 1.2.1, default parameters, Sievers et al. [2011]) and used these to reconstruct phylogenetic gene trees using RAxML (version 8.2.9, default parameters, Stamatakis [2014]). I tested the resulting trees for the monopholy of the orthogroup sequences (via DendroPy, version 4.1.0, Sukumaran and Holder [2010]) and rejected all orthogroups for which I could find a potential non-clade homologue grouping within the subtree of orthogroup sequences (see figure 2.6).

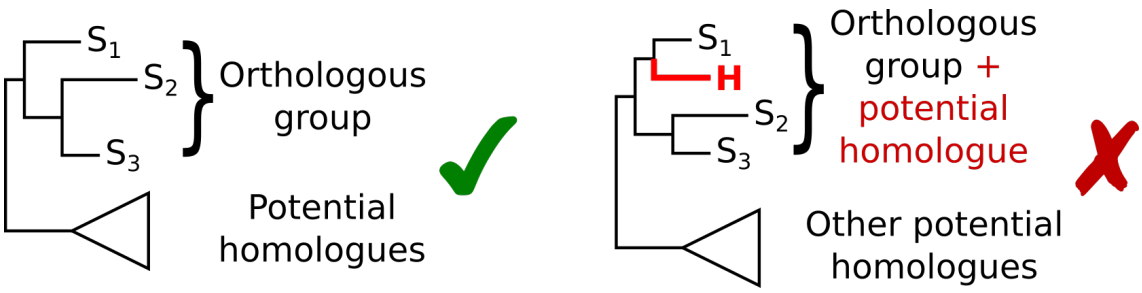


Figure 2.6.: 1st validation check to confirm a monophyletic grouping of orthogroup sequences after adding potential homologues. S_1, S_2 and S_3 are genes that have been inferred as orthologous to each other with no other orthologues outside the clade in question. The NCBI database was used to find putative homologues (including H) based on sequence similarity to the orthologous sequences. **Left**: The orthogroup forms a monophyletic clade in the reconstructed gene tree, i.e. S_1, S_2 and S_3 are closer related to each other than sequences from outside their clade. - **Right**: One or more potential non-clade homologues (H) have been inferred to diverge from within the orthologous group. The gene tree is not congruent with the species tree invalidating the orthologous group.

Using a reciprocal best bi-directional hit approach to validate clade specificity

My second validity check tests the relationship between an orthogroup and its putative homologues (implementation: `OG_rBBH_test.py`, see Appendix B.1). The previous step established, that the sequences within the remaining orthogroups are more similar to each other than to any potential outgroup homologues I found searching the NCBI RefSeq database. However, I have not yet shown that the outgroup sequences are non-orthologous to the orthogroup sequences. If there is an orthologous relationship between the orthogroup sequences and any outgroup sequences, than this would refute the inferred clade specificity of my orthogroup. The goal of my second validation step is to show that the outgroup sequences are in a paralogous relationship with the orthogroup sequences.

To prove a paralogous relation between the orthogroup and outgroup homologues I use a reciprocal bidirectional best hit (BBH) approach. BBH is defined as a pair of sequences between two species that are closest to each other compared to all other

sequences between those species. When using the sequence of the first species as a query, the sequence found in the other species will be the closest, i.e. most similar, sequence of all sequences within the second species and vice-versa. BBH approaches have been successfully used to identify orthologues (Wolf and Koonin [2012], disadvantages discussed in Dalquen and Dessimoz [2013]).

Based on the gene tree reconstructed from an orthogroup and its potential homologues I extracted the sequence(s) that grouped closest (i.e. one split away, fig. 2.7) to the orthogroup's subtree. I queried the closest grouping sequence(s) against the genes of all species represented within the orthogroup to find the most similar sequences. The return of any of the orthgroup's sequences (as opposed to a paralogous sequence within the orthogroup species) as best hit would confirm a BBH. I consider these BBH pairs to be indicative of an orthologous relation which therefore breaks the clade specificity of the orthogroup. After excluding all groups for which I was able to find BBH cases, I reach my **final set of clade specific orthologous genes**.

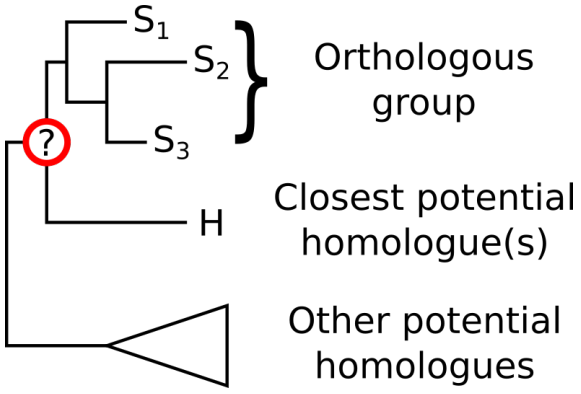


Figure 2.7.: 2nd validation check to confirm that orthogroup members are not orthologous to outgroup homologues. S_1, S_2 and S_3 are genes that have been inferred as orthologous to each other with no other orthologue outside the clade in question. Potential homologues are sequences similar to the orthogroup members found in the NCBI database, but not part of the clade of interest. H is the closest putative homologue (or set of homologues). A reciprocal best bidirectional hit approach is used to infer if the relationship between the orthogroup and H ("**?**") is orthologous. If so, H breaks the clade specificity and the orthogroup is rejected.

A final merging step was necessary due to my treatment of OrthoInspector's results. OrthoMCL and OrthoFinder both cluster putative orthologues into separate orthologous groups. Each group represents a set of orthologous genes that originated from a common ancestral gene. OrthoInspector does not provide these groupings and instead infers relationships between a gene and all its putative orthologues. Starting from different orthologous sequences, overlapping but not necessarily identical sets of genes can be found (fig. 2.3). The reason for this is that the search sensitivity depends on the sequence that is used to find putative orthologues. Throughout my pipeline, these overlapping groups were treated as separate sets of orthologous genes and subjected to the same merging steps and validity checks. Due to this I retained several overlapping groups at each step. As these groups have all passed both validity checks, I am able to combine the overlapping groups in my **final set of orthologous genes** into merged orthogroups.

2.5. Discussion

Orthology inference poses challenges both computationally and for biological interpretation. The fact that so many different programs exist to address the same issue shows us how difficult not just solving, but also understanding the problem is. Each method seems to outperform other methods under certain conditions which highlights how little we know about comprehensively retracing evolutionary changes, especially if we lack knowledge about past events or how past events relate to the data we can observe today.

Method parametrisation increases optimisation complexity. Many orthology inference methods allow the user to fine-tune parts of the analysis. As an example, the MCL program, which is part of the OrthoMCL and OrthoFinder pipelines, has a scheme parameter (`-scheme`) and an inflation parameter (`-I`), both of which affect the clustering of sequences into orthogroups. At the MCL step of the OrthoMCL pipeline, the manual only refers to the use of `-I 1.5` as an example (and presumably default) option. Without an in-depth knowledge about how MCL works and how setting these parameters affect the analysis, users have to assume that the default option is well-suited to their data and biological question. However, this default option has been chosen by the method creators to optimise the outcome based on their test data, which can be optimal to a specific question or can be a good, but not optimal, setting averaged over a range

of test cases. Even when given enough time and resources to test every potential parameter combination, we would only be able to tell which settings result in the most likely explanation of what we can observe today, but may not inform us about how these parameters relate to the underlying biology.

Heterogenous data may affect broad application of orthology prediction methods. There is an implicit assumption that orthology inference methods are capable to process a whole dataset in an optimal manner using a single set of parameters. A more sensitive setting can be useful to connect distantly related sequences or explore potential connections even at the cost of higher false positive results. More specific settings will help to decrease false positives if more conservative results are preferred. A broadly sampled dataset, however, might need different settings in specific subsets to yield accurate results. OrthoFinder is the only method used here that partially addresses data heterogeneity through its gene length normalisation. This normalisation is applied to each pairwise species comparison to reduce bias caused by phylogenetic distance between species or gene length. However, this normalisation computes an average over the whole species not taking into account that some genes may evolve faster compared to others within the same species. Akin to phylogenetic models that account for differences in nucleotide or amino acid changes between or within sequences, a "taxon- and gene-heterogenous" orthology approach might change the outcome drastically.

3. Inferring and validating clade specific orthologous groups for Bilateria, Protostomia and Deuterostomia

The goal of my orthology inference approach is not only to identify orthologous genes that are specific to a clade of interest, but also to provide a way to validate these results accounting for potential biases found in previously used approaches. In the previous chapter I described my approach to remedy potential problems introduced by the use of a specific orthology inference method and taxon sampling bias.

In this chapter I present the results of my pipeline applied to our three animal clades of interest: Bilateria, Protostomia and Deuterostomia. The results of each step highlight the potential for inferring false positives.

3.1. Results of the individual orthology inference methods - primary set of orthogroups

My first goal is to establish a primary set of orthogroups. These are the individual results of the orthology inference methods that I used choosing to keep the default parameters to avoid the introduction of additional biases. The resulting groups represent the starting point to find orthologous gene sets specific to our clades of interest.

The groups inferred from the different orthology methods were filtered according to their prevalence in the metazoan clades Bilateria, Protostomia and Deuterostomia. For each clade I filtered the orthogroups to fulfil two criteria to establish clade specificity:

exclusivity and inference in ancestral lineage. Exclusivity follows from the absence of any outgroup species (e.g. no non-bilaterian species in bilaterian specific orthogroups). Inference in a clade's ancestral lineage follows from the presence of an orthogroup's member sequences within all clades that diverged from the ancestral node: for Bilateria I required representatives in both protostomes and deuterostomes, for Protostomia I required at least one ecdysozoan and one lophotrochozoan, and for Deuterostomia at least one chordate and one ambulacrarian.

In this section I will quickly explain the differences between the three methods I used (OrthoMCL, OrthoFinder and OrthoInspector) and present the results of each with regards to the filtering for our clades of interest (table 3.1).

	OrthoMCL	OrthoFinder	OrthoInspector*
Orthogroups total	126,280	57,973	334,826
bilaterian[1]	5,748	6,967	58,702
protostome[2]	2,449	1,959	14,558
deuterostome[3]	1,666	1,607	13,503

Table 3.1.: **Primary set of clade specific orthogroups**,
Results of each orthology inference method
[1]: representative sequences in both Protostomia and Deuterostomia,
[2]: representative sequences in both Ecdysozoa and Lophotrochozoa,
[3]: representative sequences in both Ambulacraria and Chordata,
*: OrthoInspector's overlapping sets of orthologous relationships are not directly comparable with orthogroups inferred by other methods,
Inferred clade specific orthogroup do not contains any outgroup sequences.

All orthology methods use similarity scores between sequences as input to infer orthologous relationships. To generate these similarity scores I used DIAMOND (version 0.8.18.80, default parameters, Buchfink et al. [2015]) to compute scores between all pairs of sequences. The similarity scores can then be used to estimate the relatedness between the sequences.

OrthoMCL (version 2.0.9) is currently the most widely used orthology inference method (Li et al. [2003], over 2,800 citations, December 2018). OrthoMCL's pipeline includes a step to calculate pairwise gene similarity using BLAST. This step can be skipped, if similarity scores have been precalculated. These similarity scores are used to generate a graph/network in which nodes represent genes and each edge between two nodes rep-

resents their pairwise sequence similarity (fig. 3.1). Edges are weighted according to the computed similarity scores. Markov Clustering (MCL, van Dongen [2000]), a statistical method to find well connected subgraphs, is used to divide the network into distinct clusters. Each of these clusters represents a putative set of orthologous genes and recent in-paralogues (paralogues only occurring within one species) (fig. 3.2).

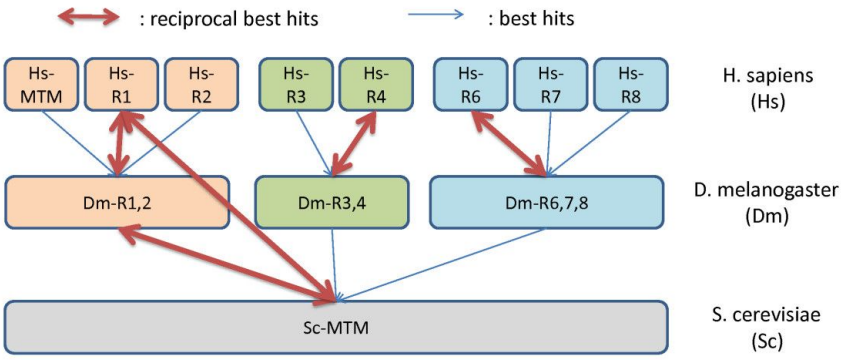


Figure 3.1.: Graph visualising similarity between genes of interest (modified from Linard et al. [2011]). Nodes represent sequences from 3 different species (right). Directed edges represent finding the most similar sequence (best hit, e.g. by using BLAST) in a different species using the edge origin as a query. This graph is used to cluster similar genes into putative orthologous groups (fig. 3.2).

OrthoMCL identified a total of 126,280 orthogroups within our data. Based on these orthogroups I inferred their representation within my clades of interest. Clade specific orthogroups contain no sequences found outside their respective clade and are inferred as present in the clade's last common ancestor. I found 6,748 groups of these to be specific to Bilateria, 2,449 groups for Protostomia and 1,666 groups for Deuterostomia.

The authors of OrthoFinder [Emms and Kelly, 2015] point out potential flaws in using BLAST scores in orthogroup detection (e.g. used by OrthoMCL). They show that short sequences fail to produce high bit scores or low *e*-values which are used to ascertain gene similarity. Using all comparisons of genes between two species, they established what they deem good representative bit scores for a given sequence length to avoid sequence length bias. They also amended the scores based on the average similarity of genes between any two species in an effort to reduce phylogenetic distance bias. These two measurements lead to similarity scores that are independent of sequence length (long vs. short orthologues) and phylogenetic distance (orthologues in closely

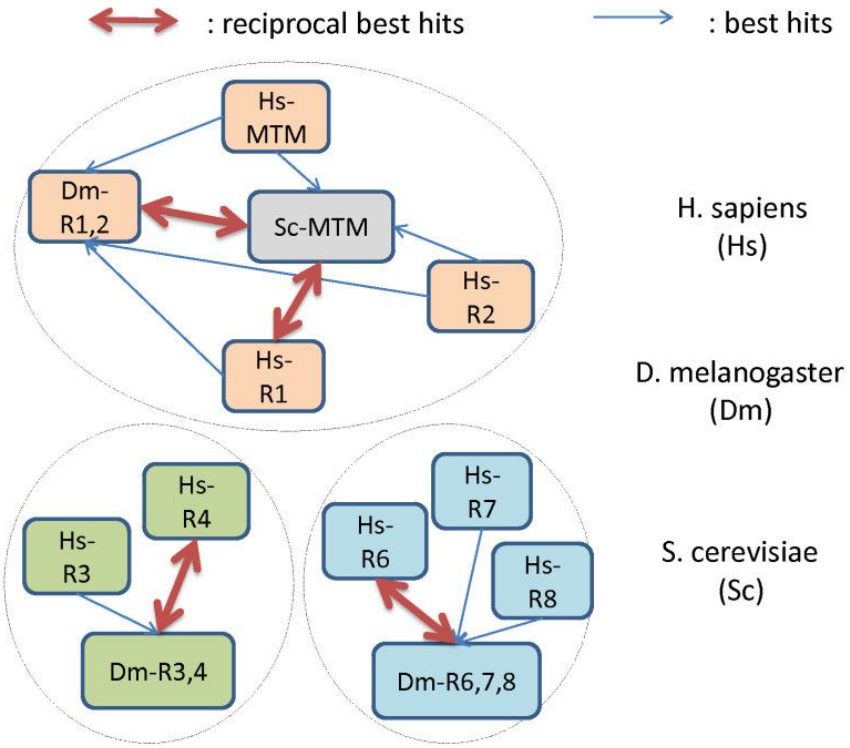


Figure 3.2.: Example clustering of putative orthologues (modified from Linard et al. [2011]) based on a graph representing the pair-wise most similar sequences between different species (fig. 3.1). Genes were grouped according the MCL algorithm which identifies well connected subgraphs. Edges between these subgraphs were pruned to form clusters of putative orthologous sequences, i.e. orthogroups. In this example, the orange and grey cluster represents the orthologous sequences present in all included species with a lineage specific duplication in humans (genes Hs-MTM, Hs-R1, Hs-R2). The green and blue clusters only exist in humans and fruit flies representing paralogous clusters that originated from a gene duplication in the ancestor of humans and fruit flies, but after the divergence from the common ancestor with *S. cerevisiae*.

vs. distantly related species). Their test results show that these amendments lead to an improvement of both recall (proportion of detected true orthologues compared to all existing orthologues) and precision (proportion of detected true orthologues amongst all inferred orthologous relations including false positives).

OrthoFinder inferred 57,973 orthologous groups within our data set, less than half as many as OrthoMCL. This trend can also be seen for the groups inferred for protostome specific orthogroups (1,959) and deuterostome specific orthogroups (1,607). Surprisingly

though, OrthoFinder yielded more orthogroups specific to Bilateria (6,967) compared to OrthoMCL. I estimated how sequences from the bilaterian specific orthogroups differ between these two methods: about 6% of the bilaterian orthogroups inferred by OrthoMCL contain sequences which are part of OrthoFinder's protostome or deuterostome specific clusters, i.e. the OrthoMCL groups were split in OrthoFinder's clustering. In the reverse direction, 24% of the OrthoFinder's bilaterian clusters are split into OrthoMCL protostome and deuterostome clusters. The higher number of bilaterian specific clusters in OrthoFinder can be explained by the combination of protostome and deuterostome clusters into larger bilaterian specific clusters.

OrthoInspector [Linard et al., 2011] has shown to be more sensitive when inferring orthologous relationships compared to OrthoMCL. Unlike OrthoMCL (and OrthoFinder), OrthoInspector does not use a graph based approach to cluster similar proteins into orthogroups. Starting with a seed protein sequence it searches for all orthologous sequences and their in-paralogues (fig. 3.3).

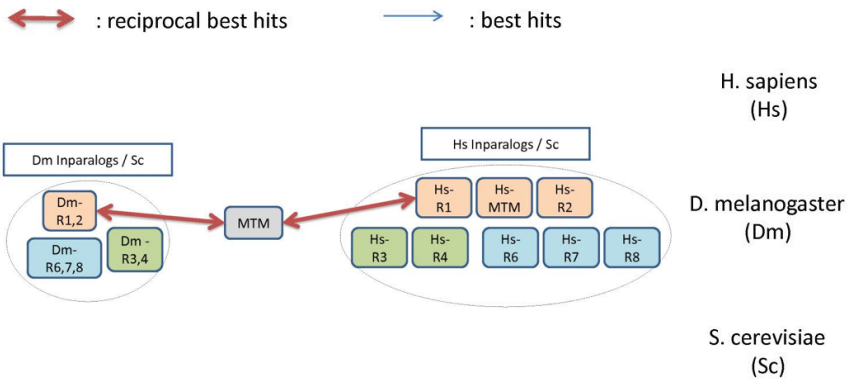


Figure 3.3.: Orthology inference using OrthoInspector (modified from Linard et al. [2011]): The myotubularin gene (MTM, grey node) of *S. cerevisiae* is used to find putative orthologous sequences in all other included species. Edges represent the best reciprocal hits found after searching for similar sequences. Not represented are the similarity scores between all other sequences used to identify the remaining putative orthologues (sequences in circles).

The outputs of OrthoInspector and OrthoMCL/OrthoFinder are not directly comparable. OrthoMCL and OrthoFinder divide the entirety of the sequence relations into distinct clusters that do not allow for overlap. OrthoInspector creates a set of orthologous relationships using each individual sequence as a starting point. Depending on

each set's starting sequence, other sets might be completely identical or just partially overlapping. To overcome the difference in output between OrthoInspector and OrthoMCL/OrthoFinder I decided to treat each unique set of orthologous relations as individual orthogroups which allows me to integrate these groups with the results of the other methods. The downside of this approach is an artificially inflated number of orthogroups inferred by OrthoInspector (total: 334,826, Bilateria: 58,702, Protostomia: 14,558, Deuterostomia: 13,503).

3.2. Merging of orthology inferences - secondary set of orthogroups

The first step of my pipeline tries to merge the different orthology inference results into orthogroups that are not affected by a single method's inherent bias(es). For the merging I consider three possibilities: i) groups identical between methods, ii) groups in partial agreement (orthogroup in one method is split into subgroups in other method(s), see detailed explanation in previous chapter), and iii) groups in disagreement. Groups of the latter category are excluded from further analysis as these represent results that are method dependent.

For Bilateria I found 450 groups to be identically inferred by all 3 methods. This represents only 6.5-7.8% of the groups inferred by OrthoMCL and OrthoFinder (primary orthogroups). After I applied the partial agreement criterion this number increased to 2,608 (37-45% of primary orthogroups).

I observed a higher congruence of inferred results for Protostomia. I found 320 protostome specific orthogroups (13-16% of primary orthogroups) to be inferred identically by all orthology inference methods and 1,014 orthogroups (41-52% of primary orthogroups) to be in partial agreement.

For Deuterostomia I observed a similar congruence comparable to protostomes . All inference methods identified the same 200 deuterostome specific orthogroups (12% of primary orthogroups) and partially agreed on 740 orthogroups (44-46%).

3.3. Validation of the secondary set of orthogroups

In this step I attempt to disprove each of the previously established putative clade specific orthogroups by searching for similar sequences in each clade's respective outgroup. I sourced these potential homologues from the NCBI RefSeq database using the secondary set of orthogroups as queries for a BLAST search (default parameters, e.g. *e*-value = 10) against the database. Afterwards I looked for a monophyletic grouping of the orthogroup sequences within a reconstructed gene tree that includes the potential homologues (monophyly test). I subjected the orthogroups that passed this step to a second validation test. The second check confirms that the closest potential homologue is **not** an orthologue, based on the reciprocal best bidirectional hit principle (paralogy test). I deem groups that pass both tests not to be affected by taxon sampling bias. I applied a final merging step to account for overlapping groups, which were caused by the previously described treatment of the OrthoInspector results. These orthogroups respresent my final set of orthogroups.

The most interesting cases are orthogroups for which I am unable to find any potential homologues (orthologues or paralogues) outside their respective clade. This implies that these groups have descended from a gene gained in the lineage of the clade's respective ancestor. Gene gain events are expected to be a rare occasion, a fact that is reflected by the small number of groups that match these criteria: I was unable to find any potential non-clade homologues for 7 bilaterian specific, 16 protostome specific and 8 deuterostome specific orthogroups (novel groups).

My monophyly and paralogy tests identified problematic cases which I rejected in all of our clades of interest (table 3.2). Bilaterian specific groups, for example, first dropped to 930 then 109 remaining orthogroups that passed both validation checks. The final merging resulted in 65 distinct groups of orthologous genes. The final set of bilaterian specific orthogroups therefore comprise 72 groups of orthologous genes. 44 protostome specific orthogroups passed both rounds of validation. Merging and adding to the previously identified novel groups I identified 51 (35+16) final orthogroups for Protostomia. I observed the lowest amount of final orthogroups for Deuterostomia. 18 orthogroups passed validation and merging totalling 24 (after including 6 novel groups) as the final set of deuterostome specific orthogroups.

Bilateria

	OrthoMCL	OrthoFinder	OrthoInspector
Orthologous Groups total	126,280	57,973	334,826
clade specific	5,748	6,967	58,702
Agreement	450		
Partial agreement	2,608		
1st validation	930 + 7		
2nd validation	109 + 7		
after final merging	65 + 7		

Protostomia

	OrthoMCL	OrthoFinder	OrthoInspector
Orthologous Groups total	126,280	57,973	334,826
clade-specific	2,449	1,959	14,558
Agreement	320		
Partial agreement	1,014		
1st validation	504 + 16		
2nd validation	44 + 16		
after final merging	35 + 16		

Deuterostomia

	OrthoMCL	OrthoFinder	OrthoInspector
Orthologous Groups total	126,280	57,973	334,826
clade-specific	1,666	1,607	13,503
Agreement	200		
Partial agreement	740		
1st validation	150 + 8		
2nd validation	21 + 8		
after final merging	18 + 6		

Table 3.2.: Numbers of inferred orthologous groups after each step of my pipeline. **Clade specific** groups must include at least one representative in both of their respective daughter clades and no sequence from outside their respective clade. Groups in **agreement** were identified identically by all 3 methods. Groups in **partial agreement** were inferred as a single orthogroup in one method, but split into several orthogroups by other methods. **Validation** steps use potential out-clade homologues found in the NCBI RefSeq database to validate the clade specificity of each group. The numbers after + represent orthogroups for which I was unable to find any non-clade homologous sequences. **Final merging** merges groups that share sequences into a single group.

3.4. Functional analysis of final set of orthogroups

For each final set of clade specific orthogroups I used InterProScan (IPS, Jones et al. [2014]) to predict protein domains to identify their potential function. IPS scans protein sequences for known protein domains. Protein domains are identified according to domains represented in the Pfam database [Finn et al., 2016]. The results are presented in tables 3.3, 3.4, 3.5, 3.6 and 3.7.

Bilaterian specific orthogroups without non-bilaterian homologues

Group ID	Pfam ID	Species	Pfam domain
G0003	PF09405	1/3	CASC3/Barentsz eIF4AIII binding
G0510		2	
G0660	PF15042	2/5	Late cornified envelope-like proline-rich protein 1
G0665		2	
G0900		2	
G1577	PF01805	1/4	Surp module
G1795		3	

Protostome specific orthogroups without non-protostome homologues

Group ID	Pfam ID	Species	Pfam domain
G0035		2	
G0098		2	
G0120		2	
G0195		2	
G0213		3	
G0292		2	
G0441		2	
G0475		2	
G0638		2	
G0775		2	
G0784		2	
G0864		3	
G0945		2	
G0984		2	

Deuterostome specific orthogroups without non-deuterostome homologues

Group ID	Pfam ID	Species	Pfam domain
G001+G161		5	
G186		2	
G272	PF01390	2/2	SEA domain
G297		4	
G494+G714	PF07557	2/3	Shugoshin C terminus
G567	PF00010	1/2	Helix-loop-helix DNA-binding domain
	PF16059	2/2	Domain of unknown function (DUF4801)

Table 3.3.: Pfam domains identified in final clade specific orthologous groups for which I was unable to find potential homologues outside their respective clade. Species states the number of species for which I identified the listed Pfam domains in their respective orthogroup sequence. Groups for which I did not find any Pfam domains only list the number of represented species.

I observed a difference in the number of protein domains identified between final orthogroups for which I was unable to identify any potential outgroup homologues and those for which I could. Table 3.3 shows that I was unable to identify Pfam domains for most of these groups. Most strikingly, I was unable to identify a single protein domain in the protostome specific orthogroups without non-protostome homologues. For Bilateria and Deuterostomia about half of the groups did not contain identifiable Pfam domains.

Group ID	Pfam ID	Species	Pfam domain
G0023+G0313+G0360+G0417+G0545+G1908	PF00096	1/7	Zinc finger, C2H2 type
	PF13894	1/7	C2H2-type zinc finger
	PF13912	1/7	C2H2-type zinc finger
G0087	PF00057	2/2	Low-density lipoprotein receptor domain class A
G0111+G0861+G2376	PF00002	6/6	7 transmembrane receptor (Secretin family)
	PF02793	5/6	Hormone receptor domain
G0133	PF00178	9/9	Ets-domain
	PF02198	8/9	Sterile alpha motif (SAM)/Pointed domain
G0172+G1131+G1151	PF00209	13/13	Sodium:neurotransmitter symporter family
G0182	PF00069	4/4	Protein kinase domain
	PF03607	4/4	Doublecortin
G0190	PF00431	4/4	CUB domain
	PF01400	4/4	Astacin (Peptidase family M12A)
	PF07645	1/4	Calcium-binding EGF domain
G0203+G0735	PF00071	14/14	Ras family
G0221+G0223+G0461+G1699	PF00104	4/12	Ligand-binding domain of nuclear hormone receptor
	PF00105	12/12	Zinc finger, C4 type (two domains)
G0328	PF07714	2/2	Protein tyrosine kinase
G0333	PF00651	8/8	BTB/POZ domain
	PF01344	7/8	Kelch motif
	PF07707	8/8	BTB And C-terminal Kelch
	PF13964	6/8	Kelch motif
G0369+G1181	PF00515	1/11	Tetratricopeptide repeat
	PF13174	1/11	Tetratricopeptide repeat
	PF13176	6/11	Tetratricopeptide repeat
	PF13181	4/11	Tetratricopeptide repeat
	PF13424	11/11	Tetratricopeptide repeat
	PF13432	1/11	Tetratricopeptide repeat
	PF14938	1/11	Soluble NSF attachment protein, SNAP
G0374	PF00168	2/2	C2 domain
G0381	PF00307	3/3	Calponin homology (CH) domain
	PF10541	1/3	Nuclear envelope localisation domain
G0387		3	
G0428	PF05380	2/2	Pao retrotransposon peptidase
G0443+G0624+G2449	PF00001	5/5	7 transmembrane receptor (rhodopsin family)
G0473	PF00400	7/7	WD domain, G-beta repeat
	PF03451	7/7	HELP motif
	PF03607	4/7	Doublecortin
	PF12894	2/7	Anaphase-promoting complex subunit 4 WD40 domain
G0512+G1108	PF05485	1/4	THAP domain
G0526	PF00665	2/2	Integrase core domain
G0527	PF00110	13/13	wnt family
G0528	PF00003	5/5	7 transmembrane sweet-taste receptor of 3 GCPR
G0543	PF00023	4/5	Ankyrin repeat
	PF07525	4/5	SOCS box
	PF12796	5/5	Ankyrin repeats (3 copies)
	PF12874	1/5	Zinc-finger of C2H2 type
G0563+G1100+G1289	PF00875	6/6	DNA photolyase
	PF03441	6/6	FAD binding domain of DNA photolyase
G0615+G0673+G0717+G0760+G1198+G1393+ G1566+G1787+G1910+G2227+G2310+G2529	PF00001	12/12	7 transmembrane receptor (rhodopsin family)
G0725+G0763+G0835+G0934+G1167+G1663+ G2218+G2503	PF00651	16/17	BTB/POZ domain
	PF01344	16/17	Kelch motif
	PF07646	1/17	Kelch motif
	PF07707	16/17	BTB And C-terminal Kelch
G0739	PF00665	2/2	Integrase core domain
G0768	PF00096	4/15	Zinc finger, C2H2 type
	PF12874	5/15	Zinc-finger of C2H2 type
	PF13894	2/15	C2H2-type zinc finger
	PF13909	10/15	C2H2-type zinc-finger domain
	PF13912	11/15	C2H2-type zinc finger
G0780	PF00096	5/8	Zinc finger, C2H2 type
	PF12874	7/8	Zinc-finger of C2H2 type
	PF13894	2/8	C2H2-type zinc finger
	PF13912	6/8	C2H2-type zinc finger
	PF16622	1/8	zinc-finger C2H2-type
G0796	PF00001	3/3	7 transmembrane receptor (rhodopsin family)
G0810	PF12937	2/2	F-box-like
	PF13516	2/2	Leucine Rich repeat

Table 3.4.: Pfam domains identified in bilaterian specific orthologous groups for which I was able to find potential non-bilaterian homologues (part 1). Species states the number of species for which I identified the listed Pfam domains in their respective orthogroup sequence. Groups for which I did not find any Pfam domains only list the number of represented species.

Group ID	Pfam ID	Species	Pfam domain
G0822	PF00071	8/8	Ras family
G0833	PF00078	1/4	Reverse transcriptase (RNA-dependent DNA polymerase)
G0834	PF00003	6/6	7 transmembrane sweet-taste receptor of 3 GCPR
G1047		2	
G1132	PF00096	4/7	Zinc finger, C2H2 type
	PF12874	6/7	Zinc-finger of C2H2 type
	PF13894	1/7	C2H2-type zinc finger
G1155	PF07679	1/3	Immunoglobulin I-set domain
	PF07686	1/3	Immunoglobulin V-set domain
	PF08205	2/3	CD80-like C2-set immunoglobulin domain
	PF13927	2/3	Immunoglobulin domain
G1164	PF00046	2/2	Homeobox domain
G1177+G1779+G1879	PF00022	4/4	Actin
	PF00646	2/4	F-box domain
G1200	PF00665	2/2	Integrase core domain
G1207	PF00001	2/2	7 transmembrane receptor (rhodopsin family)
G1223+G1920+G2283+G2400	PF00001	7/7	7 transmembrane receptor (rhodopsin family)
G1335	PF00168	3/3	C2 domain
	PF00387	3/3	Phosphatidylinositol-specific phospholipase C, Y domain
	PF00388	3/3	Phosphatidylinositol-specific phospholipase C, X domain
G1372+G1381	PF00096	1/3	Zinc finger, C2H2 type
	PF12874	2/3	Zinc-finger of C2H2 type
G1405		2	
G1487	PF00665	2/2	Integrase core domain
G1538	PF00041	2/3	Fibronectin type III domain
	PF00047	1/3	Immunoglobulin domain
	PF00102	3/3	Protein-tyrosine phosphatase
	PF01822	1/3	WSC domain
G1579	PF01433	2/2	Peptidase family M1
	PF11838	2/2	ERAP1-like C-terminal domain
G1596	PF00394	3/3	Multicopper oxidase
	PF07731	2/3	Multicopper oxidase
	PF07732	3/3	Multicopper oxidase
G1599+G1961	PF00168	11/11	C2 domain
G1631	PF00335	8/8	Tetraspanin family
G1648		2	
G2146	PF13499	5/5	EF-hand domain pair
G2163		2	
G2224	PF00008	2/2	EGF-like domain
	PF00057	1/2	Low-density lipoprotein receptor domain class A
	PF12661	2/2	Human growth factor-like EGF
G2241	PF01039	2/2	Carboxyl transferase domain
G2265	PF00400	3/11	WD domain, G-beta repeat
	PF05729	5/11	NACHT domain
	PF13271	3/11	Domain of unknown function (DUF4062)
G2330	PF01433	3/3	Peptidase family M1
	PF11838	1/3	ERAP1-like C-terminal domain
G2334	PF05380	2/2	Pao retrotransposon peptidase
G2446	PF03645	7/7	Tctex-1 family
G2476	PF00096	2/2	Zinc finger, C2H2 type
G2502	PF01184	3/7	GPR1/FUN34/yaaH family
	PF01300	7/7	Telomere recombination
G2511	PF00191	3/3	Annexin
G2522	PF00071	20/20	Ras family
	PF16474	1/20	Kinase non-catalytic C-lobe domain

Table 3.5.: Pfam domains identified in bilaterian specific orthologous groups for which I was able to find potential non-bilaterian homologues (continued). Species states the number of species for which I identified the listed Pfam domains in their respective orthogroup sequence. Groups for which I did not find any Pfam domains only list the number of represented species.

Group ID	Pfam ID	Species	Pfam domain
G0003	PF00001	8/8	7 transmembrane receptor (rhodopsin family)
G0013+G0031+G0421+G0781+G0828	PF00023	2/9	Ankyrin repeat
	PF07525	9/9	SOCS box
	PF12796	8/9	Ankyrin repeats (3 copies)
	PF13637	2/9	Ankyrin repeats (many copies)
G0045	PF02969	2/2	TATA box binding protein associated factor (TAF)
G0055+G0804	PF00096	1/6	Zinc finger, C2H2 type
	PF12874	4/6	Zinc-finger of C2H2 type
	PF13894	2/6	C2H2-type zinc finger
	PF13912	4/6	C2H2-type zinc finger
G0121+G0532	PF00105	9/9	Zinc finger, C4 type (two domains)
G0168	PF00648	3/3	Calpain family cysteine protease
	PF01067	3/3	Calpain large subunit, domain III
	PF05050	1/3	Methyltransferase FkbM domain
G0173	PF00042	2/3	Globin
G0226	PF00010	3/3	Helix-loop-helix DNA-binding domain
	PF07527	2/3	Hairy Orange
G0240	PF13520	2/2	Amino acid permease
G0256	PF00651	7/7	BTB/POZ domain
	PF01344	6/7	Kelch motif
	PF07707	6/7	BTB And C-terminal Kelch
	PF13964	2/7	Kelch motif
G0270	PF00651	3/4	BTB/POZ domain
G0312+G0664	PF01403	5/5	Sema domain
	PF01437	5/5	Plexin repeat
	PF01833	5/5	IPT/TIG domain
	PF08337	5/5	Plexin cytoplasmic RasGAP domain
G0314	PF00096	8/9	Zinc finger, C2H2 type
	PF12874	2/9	Zinc-finger of C2H2 type
	PF13912	1/9	C2H2-type zinc finger
G0319	PF00431	7/7	CUB domain
G0390	PF00112	2/2	Papain family cysteine protease
	PF08246	2/2	Cathepsin propeptide inhibitor domain (I29)
G0459	PF01344	2/2	Kelch motif
	PF13964	1/2	Kelch motif
G0496	PF00250	2/2	Forkhead domain
G0497	PF00023	1/2	Ankyrin repeat
	PF07525	2/2	SOCS box
	PF12796	1/2	Ankyrin repeats (3 copies)
G0512	PF00096	5/6	Zinc finger, C2H2 type
G0517+G0955	PF00335	6/6	Tetraspanin family
G0539		2	
G0554	PF00057	1/3	Low-density lipoprotein receptor domain class A
	PF00431	3/3	CUB domain
G0579	PF07690	1/2	Major Facilitator Superfamily
G0597	PF00096	1/5	Zinc finger, C2H2 type
	PF00651	4/5	BTB/POZ domain
	PF07707	4/5	BTB And C-terminal Kelch
	PF13894	1/5	C2H2-type zinc finger
	PF13909	1/5	C2H2-type zinc-finger domain
G0666	PF00060	7/7	Ligand-gated ion channel
	PF00497	1/7	Bacterial extracellular solute-binding proteins, family 3
	PF01094	7/7	Receptor family ligand binding region
	PF10613	7/7	Ligated ion channel L-glutamate- and glycine-binding site
G0677+G1010	PF01394	1/3	Clathrin propeller repeat
G0689	PF00230	2/2	Major intrinsic protein
G0933	PF13855	2/3	Leucine rich repeat
G0939	PF00023	1/2	Ankyrin repeat
	PF12796	2/2	Ankyrin repeats (3 copies)
G0994	PF00060	12/12	Ligand-gated ion channel
	PF00497	1/12	Bacterial extracellular solute-binding proteins, family 3
	PF10613	12/12	Ligated ion channel L-glutamate- and glycine-binding site
G1007	PF00209	2/2	Sodium:neurotransmitter symporter family
G1013	PF00023	1/3	Ankyrin repeat
	PF05033	1/3	Pre-SET motif
	PF12796	1/3	Ankyrin repeats (3 copies)

Table 3.6.: Pfam domains identified in protostome specific orthologous groups for which I was able to find potential non-protostome homologues. Species states the number of species for which I identified the listed Pfam domains in their respective orthogroup sequence. Groups for which I did not find any Pfam domains only list the number of represented species.

Group ID	Pfam ID	Species	Pfam domain
G083	PF00001	9/9	7 transmembrane receptor (rhodopsin family)
G091+G174+G634	PF00001	3/3	7 transmembrane receptor (rhodopsin family)
G110	PF04142	3/3	Nucleotide-sugar transporter
G152	PF00531	1/2	Death domain
	PF05729	2/2	NACHT domain
G203	PF00001	12/12	7 transmembrane receptor (rhodopsin family)
G358	PF00053	12/12	Laminin EGF domain
	PF00055	12/12	Laminin N-terminal (Domain VI)
	PF01759	12/12	UNC-6/NTR/C345C module
G409	PF00282	2/2	Pyridoxal-dependent decarboxylase conserved domain
G425	PF00069	2/2	Protein kinase domain
G481	PF00093	3/3	von Willebrand factor type C domain
	PF00094	3/3	von Willebrand factor type D domain
	PF01826	2/3	Trypsin Inhibitor like cysteine rich domain
	PF08742	2/3	C8 domain
G514	PF00651	6/6	BTB/POZ domain
	PF01344	6/6	Kelch motif
	PF07707	6/6	BTB And C-terminal Kelch
G586	PF00010	3/3	Helix-loop-helix DNA-binding domain
G602+G705	PF00211	4/4	Adenylate and Guanylate cyclase catalytic domain
	PF01094	2/4	Receptor family ligand binding region
	PF07701	2/4	Heme NO binding associated
	PF07714	4/4	Protein tyrosine kinase
G603	PF00282	2/2	Pyridoxal-dependent decarboxylase conserved domain
G627	PF00130	2/2	Phorbol esters/diacylglycerol binding domain (C1 domain)
G650	PF00001	2/2	7 transmembrane receptor (rhodopsin family)
G651	PF00053	9/9	Laminin EGF domain
	PF03351	1/9	DOMON domain
G672	PF00008	2/2	EGF-like domain
	PF12661	2/2	Human growth factor-like EGF
G718	PF00069	1/5	Protein kinase domain
	PF07714	4/5	Protein tyrosine kinase

Table 3.7.: Pfam domains identified in deuterostome specific orthologous groups for which I was able to find potential non-deuterostome homologues. Species states the number of species for which I identified the listed Pfam domains in their respective orthogroup sequence. Groups for which I did not find any Pfam domains only list the number of represented species.

Domains in orthogroups with potential outgroup homologues were often conserved amongst all member sequences. With few exceptions I was able to identify at least one protein domain per orthogroup present in all or almost all of the corresponding sequences. Occasionally I found domains which where either only found in a single or a small subset of member sequences.

3.5. Discussion

3.5.1. Differences between orthology inference methods

One of the goals of my pipeline is the integration of various orthology methods to prevent method inherent biases. Through my analysis of Bilateria, Protostomia and Deuterostomia I am able to demonstrate these differences (fig. 3.4). Based on the same input data, OrthoMCL (126,280 clusters) infers more than twice as many orthogroups than OrthoFinder (57,973 clusters). I found one explanation for this difference in the different average cluster sizes. OrthoMCL has a mean size of 7.1 sequences per orthogroup while OrthoFinder's mean group size is 11.8 sequences. This finding supports the reported increase in sensitivity over OrthoMCL [Emms and Kelly, 2015], which leads to the inclusion of more sequences into larger orthogroups.

Difference in method sensitivity is reflected by the different number of clusters inferred for each clade of interest. Numbers for clade specific orthogroups in Protostomia and Deuterostomia follow the general trend of OrthoMCL inferring more clusters. Contrary to that, I found more bilaterian specific clusters inferred by OrthoFinder (6,967) than OrthoMCL (5,748) even though the mean size of clusters is still higher in OrthoFinder (14.3 sequences compared to 11.1 sequences). The increased number is most likely caused by OrthoFinder's higher sensitivity that allows for a merge of clusters specific to Protostomia and Deuterostomia as inferred by OrthoMCL into larger bilaterian specific orthogroups. I compared the bilaterian groups for these two methods directly and found that 24% of the OrthoFinder bilaterian groups contained sequences that have been assigned to either protostome or deuterostome specific groups when I used OrthoMCL instead. The reverse direction is in stark contrast, with only 6% of OrthoMCL bilaterian groups getting split by OrthoFinder.

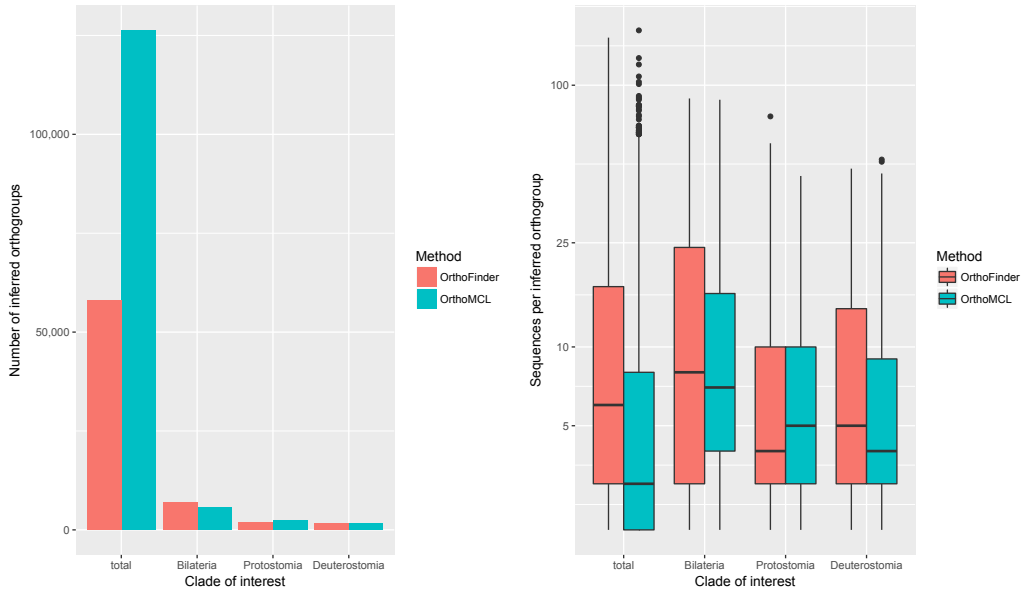


Figure 3.4.: Comparison of cluster number and sizes inferred by OrthoMCL and OrthoFinder. **Left**: The total number of clusters inferred by OrthoMCL is more than twice as many as inferred by OrthoFinder. OrthoMCL cluster numbers are higher for orthologues specific to Protostomia and Deuterostomia, but not Bilateria. **Right**: Overall, OrthoMCL creates more clusters, but with a smaller number of sequences per cluster. The distribution of cluster sizes for bilaterian, protostome and deuterostome specific orthogroups are largely overlapping.

Orthogroup numbers and sizes indicate differences in sensitivity and specificity of orthology inference methods. The increased size and decreased number of orthogroups inferred by OrthoFinder are indicators for a higher sensitivity. This increased sensitivity allows OrthoFinder to find putative orthologues missed by OrthoMCL and to include them into orthogroups. OrthoFinder is then also able to infer more connections between putative orthogroups to merge them into larger clusters that are inferred to have existed in a common ancestor and are therefore more ancient than the subclusters would have indicated. The increased sensitivity may help us find the true origin of the orthologous sequences. The downside of an increased sensitivity is the potential for false positive inference of orthology. This may cause unrelated sequences that are similar by chance to be added to orthogroups or unrelated orthogroups to be merged based on just enough similarity between their member sequences. Decreasing sensitivity, however, can lead to

the dismissal of related sequences and over-splitting of orthogroups. This trade-off led to my developing a pipeline which integrates the results of several inference methods and accepts only orthogroups that these methods agree on.

3.5.2. Merging and validation leads to rejection of most inferred orthogroups

Each step of my pipeline is designed to rigorously check the inferred orthogroups for their validity. The first step eliminates inherent biases of the orthology inference methods I used by rejecting groups that are not congruent between methods. The two validation steps eliminate taxon sampling biases by increasing the search space to a comprehensive database that includes many more species than my initial data set. My goal is the validation of orthogroups that are reliable and robust diminishing any biases.

The merging step displays the amount of disparity in resulting orthogroups caused by each method's biases. Inference congruency between methods can be shown by focusing on groups that have been identified identically. In this study I showed that the number of identical groups between methods is only a fraction of the total number of groups. 83% and more of inferred orthogroups for any given method-clade combination would be rejected if I were to apply group identity as the only criterion for method agreement. I established the partial agreement principle (fig. 2.5 B) to also include groups that may have been subject to over-splitting, i.e. a less sensitive method separated groups of genes rather than merging them into one big cluster. Using this partial agreement I added many more groups to the set of putative orthogroups inferred by all methods, decreasing the amount of rejected groups to 48% or lower with respect to the corresponding orthology inference method.

Partial agreement could lead to a higher amount of erroneously inferred orthologous relationships. Partial agreement searches the primary orthogroups for groups of one method which have been split in other methods in a way that can be remedied by merging these subsets. I do acknowledge that my partial agreement criterion harbours the risk of clustering sequences that might not be true orthologues. Without any additional checks, partial agreement will always favour the inclusion of more sensitive methods, i.e. methods that create bigger clusters. If the increased sensitivity comes at the price of an increase in false positives, partial agreement will propagate any mistakes made

during the orthology inference. I address this issue by performing validation checks on this secondary set of orthogroups to secure not only the inferred orthology, but also their clade specificity.

The first validation check led to a rejection of the majority of orthogroups from my secondary set. I search for potential non-clade homologues using the NCBI RefSeq database in order to increase the search space and decrease taxon sampling bias. I combined the sequences of an orthogroup and their potential homologues to reconstruct gene trees and check if the sequences of the orthogroup form a monophyletic clade. The monophyly criterion guarantees that no potential non-clade homologue is more similar than orthogroup sequences are to each other. If a homologous sequence groups within the clade containing the orthologous sequences, the gene tree contradicts the species tree thereby invalidating the orthology of the group. I excluded 66%/49%/79% of the secondary orthogroups specific to Bilateria/Protostomia/Deuterostomia for failing to adhere to the monophyly criterion.

The accuracy of my monophyly test depends on the reliability of the database I use to search for potential homologues. I deem NCBI RefSeq to be a well curated database of high quality to use for this purpose. Despite this, I cannot exclude the possibility of erroneously included sequences. A closer investigation of the offending homologous sequences might reveal them to be contaminations or misannotations.

Only a fraction of kept orthogroups survived the second validation check. Through the first validation check I was able to prove that all potential homologous sequences found are less similar to the orthogroup sequences than the orthologous sequences to each other. However, I have not proven those sequences to be non-orthologous. This pipeline step attempts to secure clade specificity by showing that orthogroup sequences descended from a gene that was present the clades common ancestor and are orthologous only to each other, not to sequences outside their clade. For this check I am using the reciprocal best bidirectional hit (rBBH) method to identify orthology between orthogroup members and their potential homologues. This approach detects if an orthogroup sequence and a potential homologue are closest to each other compared to all other sequences in the two corresponding species. If such a relation cannot be found than all the included non-clade sequences are at best paralogous. Only orthologous sequences would be closest to each other, as they diverge less than paralogous sequences [Koonin, 2005]. I rejected 88%/91%/86% of groups specific to Bilateria/Protostomia/Deuterostomia that passed

the first validation as I was able to identify at least one rBBH relation between sequences of these groups and their respective non-clade homologues.

I do not deny that this approach could be seen as overly conservative as I do not address problems such as differential gene loss, which could result in a rBBH relationship between paralogous sequences. This problem, known as "hidden paralogy", can cause the erroneous inference of orthology when based purely on pairwise similarity. More sophisticated methods may be able to rescue groups through the examination of the distribution and evolutionary history of paralogous copies in other species.

The exclusion of orthogroups down to a fraction of the original primary orthogroups is a result of my efforts to infer robust and validated orthogroups that are unlikely to be the result of biases. The final number of validated orthogroups matches to previous publications. For Bilateria 72 bilaterian orthogroups passed all steps of my pipeline and 85 bilaterian protein clusters were found by Krämer-Eis et al. [2016]. This shows that my pipeline is comparable in terms of quantitative outcome.

3.5.3. Little overlap of validated orthogroups with previously published findings

I looked at previous publications about bilaterian and deuterostome specific genes and compared these findings to my own analysis. Many of the published orthogroups do not match with the orthogroups inferred and validated through my pipeline. Here I will highlight a few examples that showcase how these results disagree and at what point in my pipeline these differences gained traction.

Krämer-Eis et al. [2016] report 85 protein clusters which are conserved in Bilateria. They excluded all sequences for which they could find putative non-bilaterian orthologues making the found proteins bilaterian specific. One of these bilaterian specific proteins is the transcription factor myoD which is involved in muscle development (review in Berkes and Tapscott [2005]). I looked for the corresponding sequences in my dataset and found that OrthoMCL and OrthoFinder disagreed about putative orthogroups (table 3.8). The disagreement could not be solved through my partial agreement principle, as the involved sequences belonged to more than one incongruent cluster in both methods. I did find that all genes that have been clustered together with at least one of the reported myoD sequences belong to bilaterian species. Even though I was unable to infer a large

OrthoMCL group ID	myoD sequence ID	OrthoFinder group ID
OG_12604	DME_7290	OG0001203
	DME_9683	
OG_23722	DRE_36445	
OG_24569	DRE_37465	
OG_19738	DRE_40423	OG0017120
OG_61418	CEL_3163	OG0118490
	CEL_4547	OG0119578
	CEL_22624	OG0133471

Table 3.8.: Disagreement in clustering myoD sequences. I matched the myoD sequences that were found to be part of a single cluster of orthologous sequences [Krämer-Eis et al., 2016] to my dataset (middle column). OrthoMCL and OrthoFinder produced several clusters of different size and content to group these sequences. This cluster constellation does not satisfy my partial agreement principle and was rejected at the merging step where I combine results from different orthology inference methods. DME - *D. melanogaster*, DRE - *D. rerio*, CEL - *C. elegans*

orthogroup of sequences related to myoD, my data supports the bilaterian specificity of these sequences.

[Simakov et al., 2015] reported deuterostome specific orthogroups including novelties for which they were unable to find putative homologous sequences in other clades. In total there are only 4 groups that share sequences with my findings. 3 out of these four groups contained many more sequences than my groups. I found sequences from their group 68 (annotated as "G-PROTEIN COUPLED RECEPTOR") in 2 of my final orthogroups. Group 68 contains 1,920 sequences, while my corresponding orthogroups contain only 25 sequences in total. The only exception is group 7951 ("MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE 7"). The corresponding group in my analysis contains 20 sequences, while group 7951 lists 10 sequences.

The little overlap of my results with previous findings is disconcerting. The goal of my approach was to establish and verify orthogroups for Bilateria, Protostomia and Deuterostomia. I used current orthology inference methods to establish an initial set of orthologous genes and found a large discrepancy in their results. The majority of this consensus was furthermore rejected by my validation steps. My final orthogroups are very robust to potential criticism, but do not confirm published findings. Without an

established set of orthogroups to compare to, I cannot assess, if my approach is too conservative to be meaningful. I have shown that some of the previous results should be rejected due to taxon sampling bias. These findings informed my approach to create robust orthogroups, but might increase the exclusion of valid orthogroups, i.e. increase the amount of false negatives. Further comparisons with established clusters, perhaps in other clades of interest, are necessary to establish at which step my pipeline disagrees with previous findings. Orthogroups at the evolutionary levels of Bilateria, Protostomia and Deuterostomia may need more case by case investigations to ensure the accuracy of orthology inference methods and the used thresholds.

3.5.4. Number of orthogroups in Protostomia and Deuterostomia correlates with reported differences in molecular change

One main motivation for this project was the establishment of orthologous groups specific to bilaterians, protostomes and deuterostomes. Bilaterian specific orthologues are important to reconstruct the "urbilaterian" ancestor to all Bilateria. Protostome and deuterostome specific orthogroups highlight the differences that evolved after the split from the last common ancestor to protostomes and deuterostomes, respectively. The oldest known fossil attributed to Protostomia is *Kimberella* which lived approximately 555 million years ago [Martin et al., 2000]. The oldest discovered proposed deuterostome fossil is the recently found *Saccorhytus coronarius* which dates back to 540 million years ago [Han et al., 2017]. Fossils of that time period are rare due to the soft-bodied nature of the organisms that rarely fossilise, so it is difficult to determine when the last common ancestors for Protostomia and Deuterostomia existed based on fossils alone.

Molecular data may give an insight about the time interval between the last common ancestor of all Bilateria and the last ancestors of Protostomia and Deuterostomia. Previous phylogenetic analyses showed that the amount of molecular change estimated between the bilaterian ancestor and the protostome ancestor is greater than between the bilaterian ancestor and the deuterostome ancestor [Dunn et al., 2008]. This difference could be explained by the protostome stem group existing for a longer period of time than the stem group in deuterostomes, or that there was evolutionary pressure that resulted in a higher molecular turnover.

Given a greater amount of divergence, I expected to find more protostome specific than deuterostome specific orthologous groups. The amount of genetic change should reflect the number of specific orthogroups that I can infer for each clade. Before applying my pipeline, both OrthoMCL and OrthoFinder infer more orthogroups specific to Protostomia than Deuterostomia (table 3.1). My pipeline rejected more deuterostome orthogroups (>97%) than protostome specific ones (>95%) which increased the relative difference between the two clades even further (table 3.2). I also found more than twice the number of protostome specific orthogroups without putative homologous sequences compared to deuterostome specific orthogroups (16 compared to 6). This result implies a higher number of genetic novelties that evolved in the lineage leading to protostomes.

My findings are in accordance with phylogenetic reconstructions, but require a more in-depth analysis. I found more protostome specific orthogroups than deuterostome specific ones. This supports the notion of more molecular change that occurred between the ancestor of all Bilateria and the ancestor to all Protostomia compared to the bilaterian ancestor and the deuterostome ancestor. However, this difference is only reflected in a quantitative manner. A higher number of orthogroups does not inform us about the qualitative effect these changes had. The orthogroups I found need to be further tested for their validity and function. A functional analysis is needed to confirm that the associated functions of these orthogroups are indeed present in and exclusive to their respective clade. If these functions are not exclusive to their respective clades, but the orthogroups are specific, they could be examples of non-orthologous gene displacement. Non-orthologous gene displacement describes non-orthologous genes (e.g. paralogues) that replace the original orthologues and assume their function instead [Wolf and Koonin, 2012].

3.5.5. Prevalence of orthogroup sequences

I observed a high variance in the representation of orthogroups within their respective clades. Most of the validated orthogroups comprised only 2-3 species, but some groups reached up to 80% representation (tables 3.3, 3.4, 3.5, 3.6, 3.7). Most notably, orthogroups for which I could not identify potential outgroup homologues comprised at most 5 species (table 3.3).

Orthogroups that show low representation need to be scrutinised further. For an orthogroup to be considered clade specific I required at least one member sequence of each subclade to be present, e.g. 1 ambulacrarian and 1 chordate for deuterostome specific orthogroups. For orthogroups that only consist of 2 sequences this could mean that these sequences come from 2 derived species while there is no evidence in other species of that clade. As an example, Bilaterian specific orthogroup G0655 contains only sequences from the hemichordate *S. kowalevskii* and the mollusc *L. gigantea*. Due to these species' position on the tree (fig. 2.4), a true absence of orthologues in other species would require a gene loss in many lineages such as chordates, echinoderms, ecdysozoans and many more. Another explanation could be the lack of sequencing in other species or the omission of related sequences by the orthology inference methods. A complementary search within the respective clade could identify related sequences filling this gap.

3.5.6. Functional analysis reveals clade specific orthogroups without known function

Most clade-specific groups contain known protein domains. I inferred protein domains for all protein sequences in each final orthogroup using InterPro Scan (IPS, Jones et al. [2014]). For most of the groups I was able to identify protein domains that were conserved in all or almost all member sequences supporting their relatedness (tables 3.4, 3.5, 3.6 and 3.7).

Putative *de novo* orthogroups contain no or only a few known protein domains. I found several clade specific orthogroups for which I did not find any potential outgroup homologous sequences. This implies that the ancestral genes to these groups have evolved *de novo* in the stem group of their respective clade. This makes them ideal candidates to look for clade specific changes and how these novel genes have influenced the ancestral organisms. However, for most groups I did not find any conserved Pfam protein domains (table 3.3). Only 3 of the 7 found novel bilaterian specific groups had any protein domain and these could only be identified in 1 or 2 of the member sequences. Similarly, there are only 3 novel deuterostome specific groups for which I could find protein domains, but these domains are shared by most or all of the sequences within

a group. I did not find any known protein domain for any of the 14 novel prototsome specific orthogroups.

Low representation of species in novel orthogroups is problematic. In the absence of known protein domains it is difficult to interpret function and importance of orthogroups I inferred as novel to their respective clades. Another discouraging fact is the low species representation within each novel orthogroup. I found at most 5 species to contribute to these groups, but 18 of the 27 groups are represented by only 2 species. If these species have not diverged early within their clade, it implies a large number of gene loss events that lead to the sparse representation of species within these orthogroups. The necessity of frequent loss could be an indication of a false positive finding, in which sequences have been erroneously grouped together. A more in-depth analysis of these groups is needed to gauge the importance to their clade.

4. Detection of bilaterian microRNAs in Xenacoelomorpha

Various data has been used to support a position of Xenacoelomorpha as sister to Protostomia and Deuterostomia. The first description placed them outside other bilaterian clades due to their arguably simple morphology [Westblad, 1949]. Xenacoelomorphs shared this position with Platyhelminthes as early off-shoots within the Bilateria due to their common lack of other bilaterian features, such as through-gut and body segmentation.

Molecular studied showed Platyhelminthes to be part of the Protostomia, but studies in Xenacoelomorpha are inconclusive. 18S rDNA was used to reconstruct the bilaterian phylogeny which saw platyhelminths moving to a position within the Lophotrochozoa, but leaving Acoela as sister to all other bilaterians [Ruiz-Trillo et al., 1999]. In a broad phylogenetic study, *X. bocki* instead diverged from within the deuterostomes [Dunn et al., 2008]. Philippe et al. [2011] then not only inferred the xenacoelomorphs as a monophyletic clade, but confirmed its sister relationship to the Ambulacraria (hemichordates and echinoderms).

MiRNA sequences in Acoela in comparison with other bilaterians were used to support an early branching within the Bilateria. Only 6 of the 16 [Sempere et al., 2007] or 34 [Wheeler et al., 2009] miRNAs conserved between protostomes and deuterostomes could be found in acoel species. However, sequencing of miRNAs in the acoel *Hofstenia miamia* and the xenoturbellid *X. bocki* revealed 10 and 18 more bilaterian miRNAs, respectively, to be conserved within the Xenacoelomorpha [Philippe et al., 2011]. This weakens the argument of widespread absence of bilaterian miRNAs in xenacoelomorphs, but it does not close the gap entirely.

A derived position within Deuterostomia implies a loss of bilaterian miRNA families within the Xenacoelomorpha. The reconstructed position amongst deuterostomes and the still missing absence of some bilaterian miRNA families implies a loss within the lineage leading to the xenacoelomorph ancestor. The successful sequencing of more and more bilaterian miRNAs in xenacoelomorphs have shown that this gap could be explained as an artefact caused by lack of taxon sampling or insensitivity of the sequencing approaches. I want to revisit this issue by using current sequencing methods of higher sensitivity, as well as add an additional acoel species, *Paratomella rubra*, to complement the findings about bilaterian miRNAs in xenacoelomorphs.

4.1. Introduction to microRNAs

MicroRNAs (miRNAs) are a class of non-coding RNA molecules involved in the regulation of gene expression. They are about 19-22 nucleotides long and have been found in both animals and plants. Despite their similar function in both kingdoms, miRNAs are thought to have evolved independently [Jones-Rhoades et al., 2006]. Their main mode of gene regulation involves binding to their target mRNA creating a double-strand pairing which results in a decrease in gene product through a) blocking the ribosome translating the mRNA into a protein and b) faster degradation of the mRNA (fig. 1.13)

MiRNAs are involved in many biological processes. They have been linked to different cell types, tissues and developmental stages. MiRNAs are involved in many cellular processes ranging from cell differentiation to timing developmental transitions and apoptosis [Bartel, 2004, Wienholds et al., 2005].

MiRNAs differ between plants and animals. Most of the miRNA pathway involves conserved components, such as enzymes Dicer and Dicer-like1 for cutting or the RNA-Induced Silencing Complex (RISC) which combines the mature miRNA with its target. In plants mature miRNAs match their target (near) perfectly which facilitates the hybridisation to the target mRNA transcript which blocks the protein synthesis [Jones-Rhoades et al., 2006]. In animals miRNA binding occurs through a very short sequence of 6-8 nucleotides called "seed" region which is located at the 5' end of the mature miRNA sequence (usually nucleotides 2-7). This allows for a much greater combination of interactions, where one miRNA can affect the regulation of several mRNA targets and, vice versa, one mRNA transcript may be regulated by several miRNAs.

Animal miRNAs are highly conserved. Many miRNA sequences have conservation rates of 80% or even 90% even between distantly related species. As an example, at least one homologue of the *let-7* miRNA family is perfectly conserved between humans, mice and nematodes. The high conservation rate and the potential for miRNAs to interact with many genes supports the notion of strong selective pressure. Mutations of the miRNA sequence can affect the regulation of several target genes potentially disrupting biological pathways. These effects makes strong selection against change or loss of miRNA sequences more likely.

4.1.1. Biogenesis

Mature miRNAs are not transcribed directly from the genome. At first miRNAs are transcribed as part of much longer nucleotide sequences. Several steps of enzymatic shortening and transportation from the nucleus take place, before the acting mature miRNA can be released into the cytoplasm.

In animals the primary miRNA (pri-miRNA) is the RNA sequence that has been transcribed from the genome (fig. 1.14). This pri-miRNA folds by intramolecular hydrogen bonding to contain up to six hairpin structures. These structures comprise a double-stranded stem region (which may contain imperfect pairings, i.e. bulges) and the hairpin loop. The stem region is recognised by the nuclear proteins Drosha (in vertebrates) and Pasha (in invertebrates) which cut and release the hairpin sequences, now termed precursor miRNA (pre-miRNA, fig. 4.1). RNA editing of pre-miRNAs has been reported for specific cases [Winter et al., 2009].

After release into the cytoplasm the loop part of the hairpin structure is cut off by the enzyme Dicer which results in the remainder of a double stranded sequence of about 22 nucleotides. Hairpin length and loop size are important features as they determine the efficiency of Dicer [Ha and Kim, 2014]. Even though both strands could potentially act as miRNAs, usually only one strand is functional which can be shown by a much higher abundance of the acting strand compared to the non-acting one. The active strand (the mature miRNA) joins the RNA-Induced Silencing Complex which will then enable interaction with its target mRNA(s). Unlike most genes, up to 40% of miRNA genes have been shown to reside within introns or exons of other genes [Rodriguez et al., 2004]. In this case the pre-miRNA sequences are directly spliced from the host gene and then

Figure 4.1.: Modified from Bartel [2004]: Pre-miRNAs of *lin-4* and *let-7* form a distinctive hairpin structure (double stranded stem + unpaired loop). These sequences are cut from the pri-miRNA before being exported from the nucleus. The mature miRNA sequence (red nucleotides) is later cleaved from the pre-miRNA by the enzyme Dicer.

follow the aforementioned miRNA pathway. A result of this linkage is a regulation that is tied to their host gene. In plants both cleavage steps (pri-miRNA to pre-miRNA to double stranded miRNA) occur within the nucleus and are caused by the Dicer homolog, Dicer-like1 [Jones-Rhoades et al., 2006].

4.1.2. Discovery and use as phylogenetic marker

The earliest discovery of a miRNA was that of *lin-4* in *C. elegans* in 1993 [Lee et al., 1993, Wightman et al., 1993]. But only the subsequent discovery of *let-7* and its abundance in Bilateria [Pasquinelli et al., 2000] lead to their classification as distinct regulatory elements. To classify different homologous miRNAs and to enable study of their potential targets and effects, miRNA families were introduced. These families are grouped based on their common seed sequence as this region is mostly determining which targets a miRNA will act upon and is most conserved. More than 1,500 miRNA families have been added to miRBase (release R21), a comprehensive database which collets identified mature miRNA and pre-miRNA (hairpin) sequences since it was established in 2006.

Through the discovery of miRNA family members in distantly related species it has been shown that more ancient miRNAs evolve at a slower rate than the rest of the genome, presumably due to their importance in gene regulation [Berezikov et al., 2005, Zhang et al., 2006]. Due to their role in development they have also been associated with morphological innovation leading to the emergence of more complex tissues and organ systems [Heimberg et al., 2008]. Several studies inferred that miRNAs are only gained and rarely, if ever, lost secondarily [Hertel et al., 2006, Sempere et al., 2006, Prochnik et al., 2007]. These features (slow evolutionary rate, high retention) can be very useful for phylogenetic analyses. It avoids issues with widespread missing information (e.g. due to gene loss) and systematic errors such as long branch attraction caused by high evolutionary rates in certain species. This has lead to miRNA data being incorporated in analyses to solve particularly hard cases of phylogenetic placement: a few examples include the establishment of a monophyletic clade of Cyclostomia [Heimberg et al., 2010] and Mandibulata [Rota-Stabelli et al., 2011] or the resolution of the phylogenetic relationship between Tardigrada, Onychophora and Arthropoda [Campbell et al., 2011].

Several studies have investigated miRNAs in Xenacoelomorpha with mixed results. First studies in acoel species were unable to find evidence for several miRNA families that are conserved across Protostomia and Deuterostomia [Sempere et al., 2006, 2007, Wheeler et al., 2009]. This lack of bilaterian miRNAs was seen as support for the phylogenetic position of xenacoelomorphs as sister to the remaining Bilateria. A more recent study, however, found several of the missing bilaterian miRNAs in the acoel *Hofstenia miamia* [Philippe et al., 2011]. They also added the xenoturbellid *X. bocki* to the discussion, which showed an even higher number of bilaterian miRNAs compared to its sister taxa.

More recent analyses have found evidence that may weaken the usefulness of miRNAs as phylogenetic markers. Fromm et al. [2013] have analysed flatworms (phylum Platyhelminthes) and found many otherwise conserved bilaterian miRNA families to be missing. The authors hypothesise that this is related to the simpler body plan of flatworms compared to other taxa. This would be in accordance with the hypothesis that miRNA variety drives the emergence of more complex systems [Heimberg et al., 2008]. Thomson et al. [2014] have reanalysed previous miRNA studies and found issues with naïve parsimonious approaches. Their investigation showed that loss of miRNAs is more widespread than previously believed (affecting up to 54% of miRNA families within the Bilateria). They

explain the shortcomings of simple parsimony methods given these new insights and propose the use of more sophisticated Bayesian statistical methods to estimate phylogenies.

4.1.3. Prediction, detection and validation

Originally, miRNA sequences were first identified from genomic data before being validated with the use of northern analysis. Pasquinelli et al. [2000] were the first to find conserved *let-7* sequences in a number of genomes across the Bilateria, but not in other eukaryotes. The sequences were extended and potential hairpin structures were computed. The validation of these putative miRNA sequences was time and resource intensive. RNA was isolated and separated using electrophoresis. After fixation of the RNA sequences a short labelled nucleotide sequence (probes) was added. The probe is complementary to the miRNA in question and a successful hybridisation provides evidence for the miRNAs expression in the studied organism.

Over the course of the next years several computational approaches have been developed for a more efficient identification of miRNAs. High-throughput sequencing methods such as RNA-Seq [Nagalakshmi et al., 2008] allow for the sequencing of whole transcriptomes. MiRNA identification methods scan these RNA libraries to predict conserved as well as novel miRNAs (see Bentwich et al. [2005] for an early overview).

Lai et al. [2003] developed a program called *miRseeker* which used miRNA information in closely related species (in their case, different fly species) to find homologues with little sequence divergence within the same genomic neighbourhood using BLASTN.

Lim et al. [2003] developed *MiRScan*, a program that uses machine learning to identify miRNA candidates in nematodes. After filtering the genome of *C. elegans* for regions that can form hairpins and are conserved in *C. briggsae* they were able to find 50 of the 53 previously reported nematode miRNAs [Lau et al., 2001, Lee and Ambros, 2001]. They used these 50 as a training set to identify characteristics of hairpin loops to identify true positives within their initial set of 36,000 hairpins. They reported that miRNA base pairing (i.e. small bulge sizes) in the mature miRNA region was the most important feature to identify true miRNA candidates followed by the conservation of the sequence itself. The 5' half of the sequence (which includes the seed region) was more important

than the 3' end. They experimentally validated 16 of their 35 computationally identified miRNAs.

Berezikov et al. [2005] sequenced 122 miRNA sequences from 10 primates to identify a typical miRNA conservation profile. They found high conservation in the pre-miRNA's stem region and lower conservation in the loop. Surprisingly, they found the regions flanking the pre-miRNA sequence to be most variable. They used this information to predict similar regions in human, mouse and rat as pre-miRNA candidates. All predictions were required to be found in at least 2 of these species as a form of cross validation.They identified 976 potential miRNA candidates, but were only able to experimentally validate 16 out of a set of 69 representative candidates.

Bentwich [2005] used a similar methodology to first select all potential hairpin structures from the human genome and then optimising parameters according to the characteristics of known miRNA hairpins. They found $\sim$11 million hairpins which covered 86% of the known human miRNAs, 434,239 of which passed a minimal threshold according to their algorithm *PalGrade*. They grouped the predicted miRNAs in distinct groups based on their scoring and subjected ca. 5,300 candidate miRNAs to a microarray experiment. They used a wide array of tissues to identify 886 candidate miRNAs and used their new cloning and sequencing method on 359 candidates. Using their methodology they successfully sequenced 89 new human miRNAs.

Kadri et al. [2009] developed *HHMMiR*, an approach to predict pre-miRNA hairpins without the need for conservation data. The pre-miRNA hairpin structure is divided into four regions (Loop, Extension, miRNA and Pri-extension) and each of these regions is separately modelled using a Hidden Markov Model (HMM). States within the HMM represent the amount of matches (e.g. double stranded stem region), mismatches (e.g. bulges) and indels (e.g. variable loop length) for each part of the hairpin. They combined the individual HMMs into a hierarchical HMM (HHMM) that is used to identify structures. The HHMM was trained on 527 human pre-miRNAs and $\sim$500 random hairpins. For human sequences they reached a sensitivity of up to 84% and a specificity of up to 88%. They also tested their method on other species and correctly predicted between 74.7% (mouse) and 97.4% (arabidopsis) of all hairpins tested.

Many more tools, some more specialised, some more generic to detect or predict miRNAs have been developed over the last decade [Bortolomeazzi et al., 2017].

Our biggest challenge in identifying miRNAs in Xenacoelomorpha is the lack of high quality genetic information. The aforementioned methods rely heavily on reliable prior information such as i) complete genome(s) and ii) information about closely related species. Complete genomes are needed to estimate all potential hairpin foldings and establish a positive (known pre-miRNAs) and negative (random foldings) set to train machine learning methods in distinguishing genuine sequences from false positives.

The choice of a closely related species presumes knowledge about the phylogeny of the Xenacoelomorpha and its sister taxa. The phylogenetic position of the xenacoelomorphs is a matter of contention [Telford and Copley, 2016]. The use of unsuitable species as a primer for miRNA identification tools could lead to an increase in false negatives and the interpretation of miRNAs being absent. This problem is exacerbated by the reported fast rate of evolution especially in the Acoelomorpha. MiRNAs that are affected by greater molecular change could remain undetected, following the decreasing similarity with established miRNA sequences.

Temporal and spatial specificity of miRNAs increase the difficulty of studying miRNAs in the Xenacoelomorpha. Xenacoelomorph species are known to be difficult to sample. This leads to a small number of specimens and an unavailability of developmental stages. MiRNAs are involved in embryonic development and their expression changes over time. This increases the difficulty to assess a complete set of miRNAs when using only adult specimens. We aim to compensate for that by applying sequencing methods that focus on small RNA transcripts and are more sensitive even to lower transcript numbers.

I use established knowledge about miRNA sequences to search for candidates. For my approach, I will use miRNA data that is useful independent of the Xenacoelomorpha position. I use databases about existing miRNA sequences and information about miRNA structure. I create a set of miRNA families based on the miRNA prevalence in various species. This set provides me with the information needed to search for miRNA candidates within my clade of interest. I will then use information about miRNA folding structure to evaluate these candidates. I reject candidates that are unable to form viable miRNA structures and grade viable structures according to a template. I present the candidates that match existing miRNA families and provide viable folding structures to improve our understanding of the bilaterian miRNA complement in Xenacoelomorpha.

4.2. Inference of miRNA families

4.2.1. MiRNA family characteristics

My main interest is finding similarities and differences between Xenacoelomorpha and the remaining bilaterian animals. I am mostly interested in what genetic characters are shared with other bilaterians and what bilaterian characters are absent from xenacoelomorphs. For a position of Xenacoelomorpha as sister to all other bilaterians shared characters would inform us about the common ancestor. MiRNAs shared between Bilateria and Xenacoelomorpha would confirm the existence of those miRNAs in the common ancestor. For a position of xenacoelomorphs as sister to Ambulacraria, it is important to know which miRNAs are absent. The inference of these families as conserved between protostomes and deuterostomes means that losses must have occurred in the lineage leading to the xenacoelomorph ancestor. It is important for us to understand how these losses affected the clade and what consequences the absence of specific miRNAs has.

Most miRNA detection methods try to predict any potential miRNA candidates (novel or conserved) based on transcription and genome data. Many miRNAs play key roles during the embryogenesis of animals [Bartel, 2004]. We do not possess comprehensive transcription data of high spatial and temporal resolution. This is problematic as miRNA expression levels can vary during development and during adult life. We only have the chance to sequence adult specimen, so our data will not reflect the miRNA complement that might be used mainly during development. Any miRNAs that are lowly expressed at the time of sequencing might fall below detection threshold and cause us to interpret this as an absence of the miRNA family from the species. The lack of data is mostly caused by the scarcity of the specimens and the challenges surrounding the extraction of genetic material.

Here I will be looking for specific miRNAs that have already been found in other animals rather than predicting transcribed sequences to be potential miRNAs. MiRBase [Kozomara and Griffiths-Jones, 2014] is a comprehensive database of reported miRNA findings from all domains of life. Based on published knowledge and the miRBase nomenclature I identified a set of characteristics to make miRNAs between different organisms comparable and group them into sets of miRNA families:

species - In miRBase the first 3 letters of a mature miRNA's or pre-miRNA's name denote the species for which the sequence has been identified. I use this information to infer conserved miRNA families that have presence in several clades of interest. In my analysis I focus on bilaterian specific miRNA families which by definition must be present in the genome of at least one protostome and one deuterostome to be considered.

name - The name of miRNA families is usually formed of the prefix "mir-" followed by the family's assigned number. These numbers increment whenever a new miRNA (family) has been identified. Notable exceptions are the miRNA families *lin-4*, *let-7* and *bantam* due to their establishment before naming conventions arose. Different suffixes are added to denote different loci expressing the same mature sequence (e.g. "-1", "-2" etc.) or miRNAs that are closely related (e.g. "a", "b" etc.). As I am only interested in the family itself, I ignore suffixes and include all sequences that share the same family number/name.

strand - MiRNAs reside on pre-miRNA sequences which form hairpin loop structures after being transcribed from the genome and exported from the nucleus. The hairpin loop consists of a paired stem region and an unpaired loop region. Within the stem region resides the acting mature miRNA, i.e. the sequence that will later be extracted from the pre-miRNA before it can bind to a target sequence. The acting mature miRNA sequence can reside on either the 5' or the 3' strand of the double stranded stem region. In exceptional cases both strands can be acting [Okamura et al., 2008]. In miRBase, strand information is not automatically given. The acting strand sequence is named after its family, whereas the non-acting strand is marked with an asterisk (e.g. acting miR-1 opposite non-acting miR-1*). If there is not sufficient information to determine the acting strand, the corresponding suffixes ("-5p" and "-3p") are added to the sequences' names. Information about the strand is important to group the right miRNA sequences of a given family (if both strand sequences are available) and is important to determine the position of the mature miRNA within the pre-miRNA sequence. I will use the position of the mature miRNA sequence within the pre-miRNA to extend any potential mature candidate to the full pre-miRNA sequence length while keeping the correct location of the mature candidate.

seed - MiRNAs act by hybridising to their target mRNA(s). The seed region (2nd to 8th nucleotides) is the defining character of a miRNA family which primarily determines which mRNA targets will be affected. A perfect conservation of the seed region is

necessary for a miRNA to bind to its target. A nucleotide change in the seed region will cause the miRNA to target new genes or not bind to its original target. A change in targets would alter the function of the miRNA. This makes the seed region the defining criterion of a miRNA family. The seed sequence of each family is my starting point to look for potential candidates that could fit the family's characteristics.

conservation - MiRNAs are highly conserved genetic sequences of $\sim$22 nucleotides length. I grouped all mature miRNA sequences from miRBase according to their miRNA family and their strand information. For each group I computed a multiple sequence alignment. Each group has a reference sequence, which is either a sequence from *C. elegans*, *D. melanogaster* or *H. sapiens*. If none of these species are present within the alignment, I use the first sequence extracted from miRBase as a reference sequence. I then calculated the conservation for each sequence expressed as the percentage of nucleotides which are identical between said sequence and the reference. I use the lowest conservation as threshold for the corresponding miRNA family. I reject candidate sequences that contain the seed sequence of the family, but share fewer nucleotides with the reference sequence than the conservation threshold dictates.

sequence sizes - While the length of mature miRNAs is fairly invariable (22-24 nucleotides), pre-miRNA sequences can vary in length, even within a family. I extracted the maximum lengths for both mature miRNA and pre-miRNA sequences of each family as reported in miRBase. With this information I can extract candidate miRNAs and pre-miRNAs that match other members of the family and investigate their structural composition when computing a pre-miRNA's folding structure.

4.2.2. MiRNA family inference procedure

MiRBase contains thousands of sequences from several hundred species. I created a custom Python script to scan miRBase information on mature sequences and hairpin sequences. Here I explain the general procedure my script followed in identifying miRNA families:

1. Filter by family name/number.
   Mature miRNA sequences are grouped by their number (or name) used in their description (header).

2. Filter families for clade of interest.
   Keep families that contain at least one representative sequence from each subclade, or one representative within a subclade and one outside the clade of interest (e.g. one bilaterian and one non-bilaterian sequence).

3. Separate by strand.
   Member sequences of each family were separated depending on their position on the pre-miRNA.

4. Align sequences.
   A multiple sequence aligner (Clustal Omega, Sievers et al. [2011]) is used to align grouped sequences according to their conserved residues.

5. Identify seed region.
   A sliding window of 6 nucleotides is used to identify the perfectly conserved seed region of each family. The sliding window starts at the 2nd position within the sequence alignment.

6. Compute conservation.
   The nucleotide identity between each sequence and a reference sequence (sequence from *C. elegans*, *D. melanogaster*, *H. sapiens* or first sequence is computed. The minimal conservation threshold (smallest percentage of identical nucleotides between any sequence and the reference sequence) is reported for each family.

7. Extract length information.
   The maximum length of a families member sequences is reported. For each member sequence the corresponding pre-miRNA sequence is matched. The maximum length among pre-miRNA sequences is also reported.

4.3. Bilaterian microRNA families inferred from miRBase

Bilateria is my main clade of interest to compare to Xenacoelomorpha. I used miRBase [Kozomara and Griffiths-Jones, 2014] to gain information about miRNAs found in animals

which are not xenacoelomorphs. These data I use to establish a set of families that I will compare against the data we acquired from Xenacoelomorpha.

MiRBase is a comprehensive database that collects miRNA findings across all domains of life. Its current release (R21) contains 35,828 mature miRNA sequences and 28,645 precursor hairpin sequences (pre-miRNA) from 223 species.

The first step of the miRBase scanning pipeline (see Appendix B.2) identifies miRNA families and discards miRNAs that were sequenced from non-animals. I filtered according to the description of the mature sequences stored in miRBase (file ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz). My miRNA family inference uses the naming convention of miRBase to group sequences into families. In general, miRNA families follow the same naming scheme: most families possess names that consist of the word *mir* followed by an assigned number (e.g. *mir-1*). This number increases with every newly reported miRNA sequence that could not be assigned to a previously identified family. Some families have specific names (e.g. *let-7* or *bantam*) which were assigned before the introduction of the numbering system.

The second step filtered the metazoan miRNA families for families which had to be present in the lineage leading to the last common ancestor of all Bilateria. To qualify, a family had to be present in both Protostomia and Deuterostomia or be present within Bilateria and non-bilaterians. I excluded previously discovered sequences from xenacoelomorphs as they are the focus of this investigation. I was able to identify 39 miRNA bilaterian families. All of these families were present in both Protostomia and Deuterostomia. There is no family which is present outside the Bilateria and only in either Protostomia or Deuterostomia. This indicates that there was no loss of an ancient miRNA in only one of the lineages leading to these clades' respective ancestors.

For each family I grouped the member sequences according to their position (strand) on the pre-miRNA hairpin sequence. For many sequences I was able to gain this information from the sequence description when suffixes were added to indicate strand position (e.g. cel-let-7-5p for the 5' mature miRNA sequence of *let-7* in *C. elegans*). Many sequences do not contain this information, instead they only state if the mature sequence is acting (no asterisk) or not (addition of asterisk to sequence name). In these cases I matched the mature sequence to its pre-miRNA hairpin sequence (file ftp://mirbase.org/pub/mirbase/CURRENT/hairpin.fa.gz). If the mature miRNA is part of the first half of the pre-miRNA sequence, I've assigned it to the 5' strand, if it is part of the

latter half, I've assigned it to the 3' strand. In a few cases the strands were of unequal length (mainly due to unevenly distributed bulges on the stem region), which led to the mature sequence stretching across both halves. In these cases I've manually assigned the strand position by looking up the sequence on the miRBase website.

For each family of sequences I extracted the shared seed sequence. The seed sequence is a perfectly conserved stretch of 6 nucleotides usually starting at position 2 of the mature miRNA sequence. I aligned the member sequences using Clustal Omega (1.2.4, default parameters, Sievers et al. [2011]). A sliding window of length 6 starting at the 2nd alignment site scanned across the alignment to find the potential seed region.

I've observed several issues with the miRBase data at the seed recognition step. Some families failed to contain a perfectly conserved seed sequence across the whole alignment. After manual inspection it appeared that some of the sequences may have been misattributed to their respective family. The misidentification may have been caused by less stringent classification methods in the past that used overall sequence similarity instead of seed conservation. This affected usually a very small number of sequences (1-2 sequences per family). I manually removed the offending sequences and added the seed sequence to my inference results.

Some miRNA families were no longer bilaterian specific after removing aberrant sequences. In 3 families (*mir-450*, *mir-2489* and *mir-7865*) I found member sequences that did not share the seed sequence. Removing the offending sequences would remove the only representative species of their clade (fig. 4.2). I therefore rejected these families and excluded them from the set of inferred bilaterian miRNA families.

I extracted information about the maximum mature sequence length and sequence conservation from the remaining 36 families. For the calculation of the sequence conservation I used MView (1.6.1, default parameters, Brown et al. [1998]) on each alignment of miRNA sequences (fig. 4.3). I recorded the minimum amount of conservation among all member sequence.

Lastly I needed the information about the maximum pre-miRNA sequence length in order to find candidates for precursor microRNA (pre-miRNA). For each bilaterian family I matched the mature miRNA sequences to their respective pre-miRNA sequences. I've recorded the maximum length of each family's pre-miRNA sequences as recorded in miRBase.

```
>hsa-miR-450a-1-3p MIMAT0022700 Homo sapiens miR-450a-1-3p
AUUGGGAACAUUUUGCAUGUAU--
>mmu-miR-450a-1-3p MIMAT0017182 Mus musculus miR-450a-1-3p
AUUGGGAACAUUUUGCAUAAAU--
>hsa-miR-450a-2-3p MIMAT0031074 Homo sapiens miR-450a-2-3p
AUUGGGGACAUUUUGCAUUCAU--
>mmu-miR-450a-2-3p MIMAT0004789 Mus musculus miR-450a-2-3p
AUUGGGGAUGCUUUGCAUUCAU--
>mmu-miR-450b-3p MIMAT0003512 Mus musculus miR-450b-3p
AUUGGGAACAUUUUGCAUGCAU--
>bta-miR-450a MIMAT0003834 Bos taurus miR-450a
UUUUGCGAUGUGUU-CCU-AAUAU
>hsa-miR-450b-3p MIMAT0004910 Homo sapiens miR-450b-3p
-UUGGGAUCAUUUUGCAUCCAUA-
>bta-miR-450b MIMAT0024576 Bos taurus miR-450b
UUUUGCAAUAUGUU-CCUGAAUA-
>mse-miR-450 MIMAT0024497 Manduca sexta miR-450
---GGGAUCAUUUUGCAUCCAU--
```

(a) *mir-450*: The only protostome *Manduca sexta* does not share the seed sequence.

```
>bta-miR-2489 MIMAT0011825 Bos taurus miR-2489
AAAUGACAGGGGACAUGAGUUU-----
>dme-miR-2489-3p MIMAT0012196 Drosophila melanogaster miR-2489-3p
-UAUUGUAUGU----UGUAUUUGCAGU
>dsi-miR-2489 MIMAT0012510 Drosophila simulans miR-2489
-UAUUGUAUGU----UGUAUUUGCAGU
```

(b) *mir-2489*: The only deuterostome *Bos taurus* does not share the seed sequence.

```
>prd-miR-7865-3p MIMAT0030453 Panagrellus redivivus miR-7865-3p
CUAAAUCAUGUCGAUCUGGUC----
>bta-miR-7865 MIMAT0030450 Bos taurus miR-7865
------CAGGGAGGGCAGGGGAGGG

>prd-miR-7865-5p MIMAT0030452 Panagrellus redivivus miR-7865-5p
---CGAGAUCGUUUUGAUUUUAGCG
>bta-miR-7865 MIMAT0030450 Bos taurus miR-7865
CAGGGAGGGC------AGGGGAGGG
```

(c) *mir-7865*: There is no information about the acting strand in *Panagrellus redivivus*, but both possibilities fail to align.

Figure 4.2.: Families excluded from the set of bilaterian microRNA families. Families were identified from miRBase by their common name. The families listed here failed to provide a commonly shared seed sequence (red boxes) between all sequences. The exclusion of sequences without the seed sequence removes a representative needed to infer a conservation across Bilateria.

```
Reference sequence (1): cel-miR-1-3p
Identities normalised by aligned length.

 1 cel-miR-1-3p  MIMAT0000003 Caenorhabditis... 100.0%  UGGAAUGUAAAGAAGUAUGUA-
 2 dme-miR-1-3p  MIMAT0000105 Drosophila mel...  90.9%  UGGAAUGUAAAGAAGUAUGGAG
 3 mmu-miR-1a-3p MIMAT0000123 Mus musculus m...  95.5%  UGGAAUGUAAAGAAGUAUGUAU
 4 hsa-miR-1-3p  MIMAT0000416 Homo sapiens m...  95.5%  UGGAAUGUAAAGAAGUAUGUAU
 5 gga-miR-1a-3p MIMAT0001127 Gallus gallus ... 100.0%  UGGAAUGUAAAGAAGUAUGUA-
 6 gga-miR-1b-3p MIMAT0001175 Gallus gallus ...  95.2%  UGGAAUGUUAAGAAGUAUGUA-
 7 bmo-miR-1a-3p MIMAT0004191 Bombyx mori mi...  90.9%  UGGAAUGUAAAGAAGUAUGGAG
 8 sme-miR-1a-3p MIMAT0003960 Schmidtea medi...  77.3%  UGGAAUGUCGAGAAAUAUGCAU
 9 sme-miR-1b-3p MIMAT0003961 Schmidtea medi...  68.2%  UGGAAUGUCGUGAAUUAUGGUC
10 sme-miR-1c-3p MIMAT0003962 Schmidtea medi...  66.7%  UGGAAUGUUGUGAAUAGUGUC-
11 sko-miR-1-3p  MIMAT0009607 Saccoglossus k...  90.9%  UGGAAUGUAAUGAAGUAUGUAU
12 cin-miR-1-3p  MIMAT0016393 Ciona intestin...  86.4%  UGGAAUGUAAAGAAGUAUGCGU
13 prd-miR-1-3p  MIMAT0030721 Panagrellus re...  95.5%  UGGAAUGUAAAGAAGUAUGUAG
14 gga-miR-1c    MIMAT0007503 Gallus gallus ...  86.4%  UGGAAUGGAAAGCAGUAUGUAU
15 bta-miR-1     MIMAT0009214 Bos taurus miR-1   95.5%  UGGAAUGUAAAGAAGUAUGUAU
16 cte-miR-1     MIMAT0009502 Capitella tele...  95.5%  UGGAAUGUAAAGAAGUAUGUAG
17 lgi-miR-1     MIMAT0009557 Lottia gigante...  95.5%  UGGAAUGUAAAGAAGUAUGUAU
18 spu-miR-1     MIMAT0009650 Strongylocentr...  95.5%  UGGAAUGUAAAGAAGUAUGUAU

MView 1.61, Copyright (C) 1997-2016 Nigel P. Brown
```

Figure 4.3.: I use MView to calculate the conservation threshold for each miRNA family. The lowest conservation within the *mir-1* family is 66.7% (`sme-miR-1c-3p`). This threshold is used to find potential *mir-1* candidate sequences.

I exported all relevant information to a comma separated value (CSV) file to be used with the detection pipeline. The final results of my inference of bilaterian miRNA families including relevant information for candidate identifiifaction are shown in table 4.1.

Name	Strand	Seed	$\text{Identity}_{\text{min}}$	$\text{Identity}_{\text{avg}}$	l_{miRNA}	$l_{\text{pre-miRNA}}$
let-7	5'	GAGGUA	0.636	0.901	24	122
mir-1	3'	GGAAUG	0.667	0.901	22	96
mir-7	5'	GGAAGA	0.727	0.944	24	125
mir-9	5'	CUUUGG	0.727	0.918	24	108
mir-10	5'	ACCCUG	0.625	0.886	24	110
mir-29	3'	AGCACC	0.682	0.869	23	101
mir-31	5'	GGCAAG	0.773	0.900	23	127
mir-33	5'	UGCAUU	0.857	0.960	22	101
mir-34	5'	GGCAGU	0.792	0.882	24	119
mir-71	5'	GAAAGA	0.864	0.938	23	119
mir-78	3'	GGAGGC	0.762	0.881	21	99
mir-92	3'	AUUGCA	0.500	0.876	23	101
mir-96	5'	UUGGCA	0.565	0.790	24	108
mir-100	5'	ACCCGU	0.909	0.976	24	142
mir-124	3'	AAGGCA	0.591	0.892	23	101
mir-125	5'	CCCUGA	0.591	0.866	24	110
mir-133	3'	UUGGUC	0.739	0.945	23	119
mir-137	3'	AUUGCU	0.864	0.941	23	102
mir-153	3'	UGCAUA	0.955	0.985	22	102
mir-182	5'	UUGGCA	0.750	0.872	25	111
mir-183	3'	AUGGCA	0.826	0.930	23	101
mir-184	3'	GGACGG	0.667	0.920	24	102
mir-190	5'	GAUAUG	0.654	0.832	24	101
mir-193	3'	ACUGGC	0.682	0.842	24	108
mir-210	3'	UGUGCG	0.682	0.861	23	110
mir-216	5'	AAUCUC	0.560	0.846	25	110
mir-219	5'	GAUUGU	0.636	0.908	23	102
mir-242	5'	UGCGUA	0.522	0.622	22	107
mir-252	5'	UAAGUA	0.762	0.883	23	101
mir-278	3'	CGGUGG	0.700	0.899	22	101
mir-281	3'	GUCAUG	0.789	0.886	23	102
mir-315	5'	UUUGAU	0.826	0.948	23	104
mir-365	3'	AAUGCC	0.682	0.922	22	111
mir-375	3'	UUGUUC	0.739	0.918	23	101
mir-981	3'	UCGUUG	0.870	0.934	23	97
mir-2001	5'	UGUGAC	0.909	0.966	23	77

Table 4.1.: Families identified as ancestral to Bilateria based on sequences found in miRBase including extracted information about sequence similarity of member sequences, lengths of mature and pre-miRNA sequences.

4.4. Detection of specific microRNA families from genome and transcript data

4.4.1. Detection of mature miRNA candidates from small RNA transcripts

Mature miRNAs act by hybridising with their mRNA targets within the cytosol of the cell. MiRNA detection pipelines use small RNA transcriptomic data to search for potential mature miRNA sequences (22-24 nucleotides). Our collaborator, Peter Sarkies, used RNA extracted from *Xenoturbella bocki* to sequence small (50 nucleotides or shorter) RNA sequences.

I have created a pipeline to detect mature miRNA candidates in this set of small RNA sequences based on a set of miRNA families I extracted from miRBase (implementation: `microRNA_detection.py`, see Appendix B.2). The miRNA families in question are miRNA sequences conserved across Bilateria. The results of my pipeline allow me to assess which bilaterian miRNAs are absent from the set of small transcripts and for which bilaterian miRNAs I can find viable sequences. The length of mature miRNA sequences is between 22 and 24 nucleotides. My pipeline ignores all sequences with fewer than 15 nucleotides. At this length, I am unable to conclude if the sequence is part of a miRNA or part of another sequence that got transcribed. The short sequence length increases the chance for a false positive detection as a mature miRNA candidate.

My pipeline first scans all small RNA transcripts for the occurrence of miRNA family seed sequences. These seed sequences are unique to each family and always perfectly conserved. The seed sequence determines a miRNA's target sequences which get silenced after hybridisation. It is perfectly conserved, because even small nucleotide changes affect the hybridisation to all potential targets and could introduce new targets. The seed sequence is 6-7 nucleotides long and usually starts at the 2nd nucleotide. Due to this fixed location within the mature miRNA sequence, I restrict the search area to find a perfectly matching seed sequence to the first 10 nucleotides for each small RNA transcript. Small RNA transcripts that match at least one seed sequence of a bilaterian miRNA family are kept as mature miRNA candidates.

I reject all candidates that show little conservation of the overall sequence. Each candidate so far has been identified as a potential member of a miRNA family based on the

seed sequence shared. For each family I previously established a conservation threshold based on the miRBase data. This conservation threshold (expressed as percentage of nucleotides identical between member sequences) is the smallest amount of conservation between any of the member sequences of a family and the family's reference sequence. The reference sequence of a family refers to a member sequence from a model organism (*C. elegans*, *D. melanogaster* or *H. sapiens*) or the first sequence extracted from miRBase. For each candidate I calculate the amount of nucleotides identical between the candidate and the reference sequence. I reject the candidate sequence if this value is below the minimal conservation within the family.

All found candidates are viable as mature miRNAs based on the existence of a miRNA family's seed sequence and overall sequence similarity to the member sequences of the corresponding miRNA family. The perfectly conserved seed sequence is required so that the candidate can hybridise with the same target sequences. The overall similarity to the miRNA family supports the fact that the candidate provides the same function as member sequences. A lower similarity than found could be possible, but it is not clear how this may negatively influence the target binding. I use the conservation threshold to avoid false positive detection of sequences that share the seed sequence by chance, but do not display any similarity otherwise.

A mature miRNA sequence is part of a longer pre-miRNA sequence that follows well defined steps through the miRNA biogenesis pathway. For each mature miRNA candidate I have to show that I can find a pre-miRNA candidate containing the mature candidate. This pre-miRNA candidate must also be able to fold into a specific hairpin structure. This structure is necessary for the processing steps between export of the pre-miRNA from the nucleus and the release of the mature miRNA into the cytoplasm.

4.4.2. Detection of pre-miRNA candidates based on mature miRNA candidates

Evidence for mature miRNA sequences must exist within the genome. Mature miRNA sequences must be coded for in the genome so they can be transcribed into RNA. For each mature miRNA candidate I scan the genome and its reverse complement for the putative miRNA sequence using simple text matching. I reject all mature miRNA candidates for which I am unable to find a perfect match. I require a perfect match, because it is not

trivial to ascertain if even small differences are due to an RNA sequencing error or due to a true absence of the candidate from the genome. I chose this conservative approach to reduce the number of false positive candidates that might erroneously match to the genome. Another reason for this matching step to fail is the incompleteness of the genome. My results are based on our most recent draft genomes. Improving the quality and completeness of these genomes may result in a higher recall of truly present miRNA sequences in the future.

A mature miRNA is not transcribed directly from the genome. Instead, the mature miRNA is embedded in a much longer sequence called precursor miRNA (pre-miRNA). This sequence is then processed into the short acting mature miRNA through several steps of the miRNA biogenesis (fig. 1.14). For each of my mature miRNA candidates I also have the information about location on the pre-miRNA and pre-miRNA sequence length from miRBase according to the candidate's putative miRNA family. I extract the pre-miRNA candidates by extending from the mature miRNA sequence in the genome to reach the stated length. Part of this putative pre-miRNA is a 10 nucleotide buffer between the mature miRNA candidate and the closest end of the pre-miRNA sequences. The location of the buffer depends on the mature miRNA sequence residing close to the 5' or 3' end of the pre-miRNA. This buffer sequence is needed to infer folding structures with a paired stem region which starts ahead of the mature miRNA sequence. In early attempts, I used a smaller overhang of 5 nucleotides. I tested this setting on published pre-miRNA sequences by producing RNA foldings. I noticed that several RNA folding structures would not form proper hairpins and would be rejected in the later steps of my pipeline. The increase to 10 nucleotides solved this issue.

4.4.3. Evaluation of hairpin structures from pre-miRNA candidates

Pre-miRNA sequences must be able to form a hairpin structure to be processed into mature miRNAs. After the pre-miRNA sequence has been cut from a larger transcript within the nucleus, it forms a hairpin structure (fig. 1.14). The hairpin structure comprises a paired stem region, where the pre-miRNA hybridises with itself, and an unpaired loop region. The formation of this hairpin structure is necessary for the enzyme Dicer to

attach, cut and release the double stranded stem region containing the mature miRNA sequence.

Pre-miRNA candidates that do not form hairpin structures cannot be processed according to the miRNA biogenesis pathway. I examine the ability of each pre-miRNA candidate to fold examining the most likely RNA folding structure predicted by RNAfold (version 2.4.3, default parameters, Lorenz et al. [2011]).

Each 2D RNA structure is graded according to its compliance with an "ideal hairpin". We chose the candidate pre-miRNA of *mir-34* found in *X. bocki* as a template. The *mir-34* mature miRNA candidate had a 100% conservation rate compared to the *mir-34* sequence in *C. elegans* and the hairpin structure of its potential pre-miRNA sequence (fig. 4.4) was approved by our collaborator.

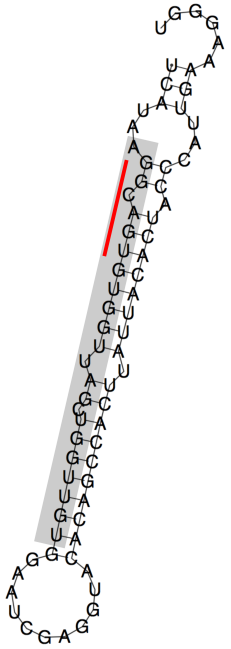


Figure 4.4.: Pre-miRNA candidate of *X. bocki* for *mir-34*: The mature candidate found in small RNA transcripts is 100% identical to the reference sequence of the *mir-34* family. Its pre-miRNA sequence was extracted from the genome and the calculated hairpin structure was approved by Peter Sarkies for its viability. I use the characteristics (e.g. bulges and loop size) of this structure as a template to evaluate other potential pre-miRNA structures. Red underlined - seed sequence of *mir-34* family, grey box - mature miRNA candidate (identical to *mir-34* sequence from *C. elegans*).

My grading algorithm first ascertains whether the predicted RNA folding structure resembles a hairpin structure at all. Hairpins must contain a stem region, a double stranded region where the RNA hybridises with itself, and an unpaired loop region. Dicer is necessary to cleave the pre-miRNA and requires a hairpin structure as well as a correct placement of the mature miRNA within said structure. I reject a candidate, if the predicted RNA structure does not form a hairpin or if the mature miRNA candidate is positioned in a way that Dicer would split the sequence. If a candidate passes this step, I assign a numerical value ("grade") to express how much certain features deviate from our ideal template miRNA candidate (see above). A numerical value of "1" is given to all pre-miRNA candidates whose parameters stay within the specified limits. I assign higher values to candidates with features that are beyond the stated limits. This does not mean that the pre-miRNA cannot be processed only that processing efficiency is more likely to be lower in these candidates than in an ideal structure. I am currently using 4 parameters to reject or grade miRNA candidates:

arms - I reject a pre-miRNA candidate if the RNA structure folds into more than one arm/loop or has no stem region. Dicer cleaves the loop region and would leave a branched structure. If no stem region is calculated (i.e. structure is circular), Dicer could not accurately recognise and cut the mature miRNA from the pre-miRNA. Both of these cases would not lead to the miRNA duplex structure which consists only of the mature miRNA and it's (imperfectly) complementing reverse strand.

position of mature miRNA - I reject a candidate if mature miRNA candidate sequence is not part of the stem region. Dicer attaches at the loop and shortly after the start of the stem region. This would lead to Dicer cutting through the mature miRNA, if it were either part of the loop structure or if the double stranded stem region were to start after the beginning of the mature miRNA sequence.

bulge size - I grade structures according to their maximum number of consecutively unpaired nucleotides in the stem region. An increased size of bulges in the stem region has been associated with a decreased efficiency of Dicer [Ha and Kim, 2014]. We have chosen the miRNA candidate of *mir-34* (fig. 4.4) as the template for scoring. Its predicted folding structure contains unpaired stem regions of at most 3 nucleotides within any given bulge. I increment a candidate's grade by 1 for each nucleotide above 3 for the largest bulge size.

loop size - I grade structures according to their loop size. Akin to large bulges, overly large loop structures decrease the processing efficiency of Dicer [Ha and Kim, 2014]. Loop size has been shown to be more variable than the stem region [Lim et al., 2003, Berezikov et al., 2005]. This allows for a greater deviation of the loop size compared to bulge sizes without compromising viability. Based on the *mir-34* candidate's loop size, I increment the grade by 1 for every 2 extra nucleotides above 11 unpaired nucleotides.

The length of pre-miRNA sequences is unknown in Xenacoelomorpha. Due to the lack of comparable species, we currently have no insight about the actual length of the pre-miRNA sequences in our clade of interest. The lack of available transcript data prevents us from accurately reconstructing the actual hairpin sequence.

My pipeline scans for a best graded, viable, pre-miRNA candidate that adheres to a family's maximum hairpin length. My pipeline incrementally shortens the pre-miRNA candidate to evaluate shorter possible hairpins. I report the best graded of these structures and its 2D folding structure as the best pre-miRNA candidate for the corresponding mature miRNA sequence.

In summary, my detection pipeline uses short RNA transcript data and data about known miRNA families to scan for mature miRNA candidates. It confirms these sequences' existence within our species by searching the genome for potential pre-miRNA sequences. It validates these findings by assessing the viability of the pre-miRNA sequence to form a proper hairpin structure which is necessary for being processed via the miRNA biogenesis pathway.

4.5. Bilaterian microRNAs in Xenacoelomorpha

4.5.1. Previous microRNA findings regarding Xenacoelomorpha

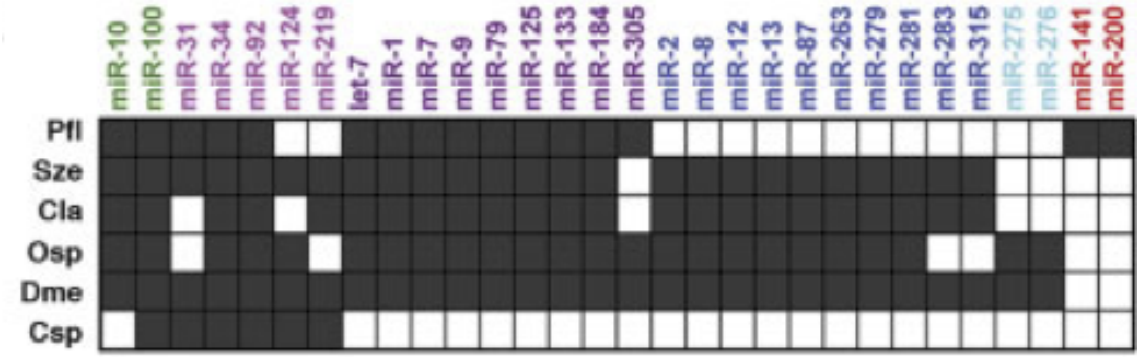
The enigmatic relationship of Xenacoelomorpha to other animals led to a search for phylogenetic characters to identify their position in the tree of life. Their apparently simple morphology [Nielsen, 1995] has been interpreted as a "primitive" state within Bilateria, a description shared with Platyhelminthes [Westblad, 1949]. Both clades were considered sister to all other bilaterians (protostomes and deuterostomes). More recent phylogenies based on molecular data revealed the platyhelminths to be a morphologically

simple but derived clade diverging from within the Lophotrochozoa [Balavoine, 1997, Carranza et al., 1997]. The position of Xenacoelomorpha has alternated between being a sister to all Bilateria and being sister to the Ambulacraria, diverging from within the Bilateria [Telford and Copley, 2016].

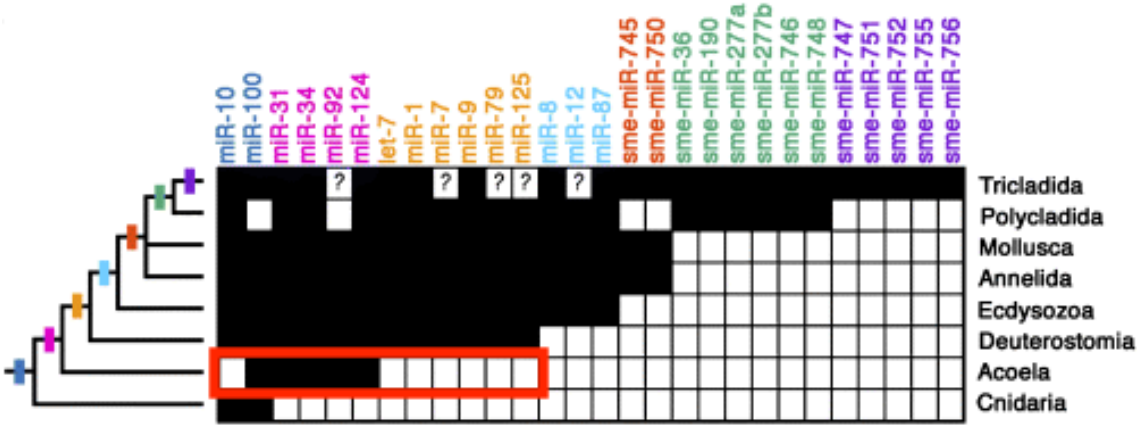
Absence of the *let-7* miRNA in acoelomorphs was seen as evidence to support a sister relationship of Xenacoelomorpha and the remaining Bilateria. Pasquinelli et al. [2000] discovered the prevalence of *let-7* amongst many bilaterians but not in Cnidaria, Ctenophora and Porifera. In 2003 they extended their search to other metazoans including the acoels *Convoluta convoluta*, *Symsagittifera roscoffensis* and *Amphiscolops* sp. [Pasquinelli et al., 2003]. They were unable to detect *let-7* in any of the acoels, an observation shared with all non-bilaterians tested. They interpret this as supporting the divergence of Acoelomorpha before the evolution of *let-7* in the lineage leading to the protostome-deuterostome ancestor.

Investigations of several bilaterian miRNAs have shown a reduced set in Acoelomorpha compared to other bilaterians. Sempere et al. [2006] used published miRNA sequence data to infer miRNA families that are conserved across the Bilateria. Their search resulted in 20 miRNAs to be conserved between protostomes and deuterostomes. They also inferred several families restricted to either protostomes or deuterostomes. They tested these miRNA families to find evidence in other metazoans including a cnidarian and the acoel *Childia* sp. 16 bilaterian miRNA families were tested, but only 6 of these could be detected in the acoel worm (fig. 4.5a). In comparison, they found all but one bilaterian miRNA and all protostome specific miRNAs in the polyclad flatworm *Stylochus zebra*. A follow-up investigation added the acoel *Symsagittifera roscoffensis* and repeated the validation of bilaterian specific miRNAs [Sempere et al., 2007]. Their results matched the previous findings leaving acoels bereft of most bilaterian miRNAs (fig. 4.5b).

A more recent analyses of genetic data shows prevalence of more miRNA families in Acoelomorpha and *Xenoturbella bocki*. The phylogenetic inference of Philippe et al. [2011] supported the previously inferred grouping of *X. bocki* and the Acoelomorpha [Hejnol et al., 2009], proposing the name Xenacoelomorpha for the unified clade. As part of their analyses they also sequenced miRNAs from another acoel, *Hofstenia miamia*, and *Xenoturbella bocki*. In *H. miamia* they found 10 bilaterian miRNAs which were not sequenced in the previous analyses involving acoels. Additionally they were able to sequence



(a) MiRNA complement of acoel *Childia* sp. (Csp, bottom row) in comparison with other bilaterian species, modified from Sempere et al. [2006];
Pfl - *Ptychodera flava*, Sze - *Stylochus zebra*, Cla - *Cerebratulus lacteus*,
Osp - *Orthoporus* sp., Dme - *Drosophila melanogaster*.



(b) MiRNA complement of *Symsagittifea roscoffensis* (Acoela, red box) and other eumetazoan clades, modified from Sempere et al. [2007].

Figure 4.5.: Presence (black) and absence (white) of miRNAs in several bilaterian species and cnidarians show a smaller miRNA complement in acoel species.

another 8 bilaterian miRNAs in *X. bocki*, but not all of the bilaterian miRNAs tested. In both species they also found the deuterostome specific miRNA *mir-103/107/2013*. They applied parsimony analysis to the absence and presence of all tested miRNAs and found narrow support of xenacoelomorphs diverging before the split of Protostomia and Deuterostomia. However, their phylogenetic inference using 197 genes placed Xenacoelomorpha within the Deuterostomia as sister to the Ambulacraria and *X. bocki* is shown to have a lower evolutionary rate than Acoelomorpha. If this phylogenetic position is correct then the the lack of bilaterian miRNAs in Xenacoelomorpha and even

more in Acoelomorpha must have been caused by losses in the lineages leading to the clades' respective ancestors (fig. 4.6).

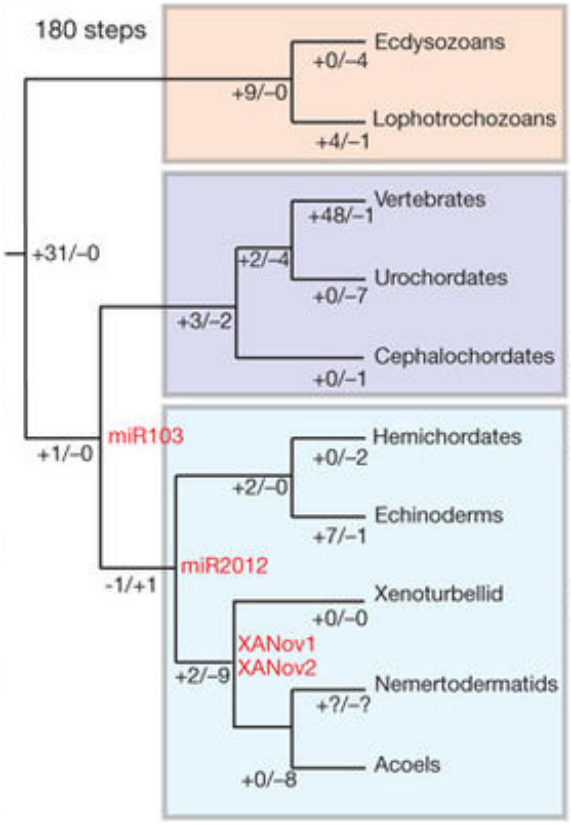


Figure 4.6.: Gains (+) and losses (-) of miRNA families based on the placement of Xenacoelomorpha as sister to Ambulacraria as inferred by Philippe et al. [2011], red - miRNAs specific to Deuterostomia (*mir-103*), Ambulacraria (*mir-2012*) and Xenacoelomorpha (*XANov1*, *XANov2*).

My goal was to re-examine the reported absence of bilaterian miRNAs in Xenacoelomorpha. Current sequencing protocols and technology allow for a more sensitive sequencing of short RNA molecules. I wanted to check, if the reported lack of bilaterian miRNAs was due to the use of less sensitive methods or if previous results can be confirmed through the use of current methodology. I wanted to add more acoel species to this ongoing investigation, including the slow evolving *Paratomella rubra* to see if the absence of many bilaterian miRNAs in previously studied acoels is characteristic for the whole clade. I also use newly assembled and improved draft genomes of xenacoelomorph species to verify my findings. My results will improve our understanding about the

presence and absence of bilaterian miRNAs within Xenacoelomorpha compared to other bilaterians.

4.5.2. Preparation and RNA extraction

My main miRNA detection pipeline requires small RNA data as input to identify potential mature miRNA candidates. We intended to sequence RNA from 3 species of xenacoelomorphs: the xenoturbellid *Xenoturbella bocki* and two acoels, *Paratomella rubra* and *Symsagittifera roscoffensis*.

RNA transcripts from *X. bocki* were extracted and then specifically size selected to retain only small RNA transcripts before being sequenced. We collected *X. bocki* off the west coast of Sweden by dredging the muddy bottom of the fjord near the Sven Lovén Centre in Fiskebäckskil. We placed one whole specimen of *Xenoturbella bocki* in a tube and removed sea water as much as possible before adding TRIzol. Tube contents were pipetted up and down to dissolve tissue. Peter Sarkies (collaborator, Imperial College London) precipitated RNA with glycogen overnight at -20°C. He extracted 2.5µg of RNA and prepared the subsequent small RNA library following the Illumina small RNA kit protocol.

The small RNA dataset for *X. bocki* consists of 4,334,980 different short sequences of length 5 to 50 nucleotides. The number of times a given unique sequence has been sequenced ranges from 1 to 621,877 (fig. 4.7).

RNA extraction from *P. rubra* failed due to low RNA amount. Helen Robertson, Fraser Simpson and I collected sand from Filey Bay, Yorkshire and extracted worms in our laboratory. We pooled ∼100 specimens removing as much sea water as possible before freezing. Unfortunately, Peter Sarkies was unable to extract enough RNA material to successfully apply the small RNA sequencing protocol.

RNA from our *S. roscoffensis* specimens has not been extracted as of writing this thesis. We collected *S. roscoffensis* from the French north coast near Roscoff. Helen Robertson and I pooled an estimated amount of 800-1000 specimens in two tubes before removing sea water and freezing. These specimens are still awaiting being processed by Peter Sarkies.

Kevin J. Peterson kindly provided me with the RNA transcript data from *S. roscoffensis* (in personal communication, data not publicly available) which was used in their paper

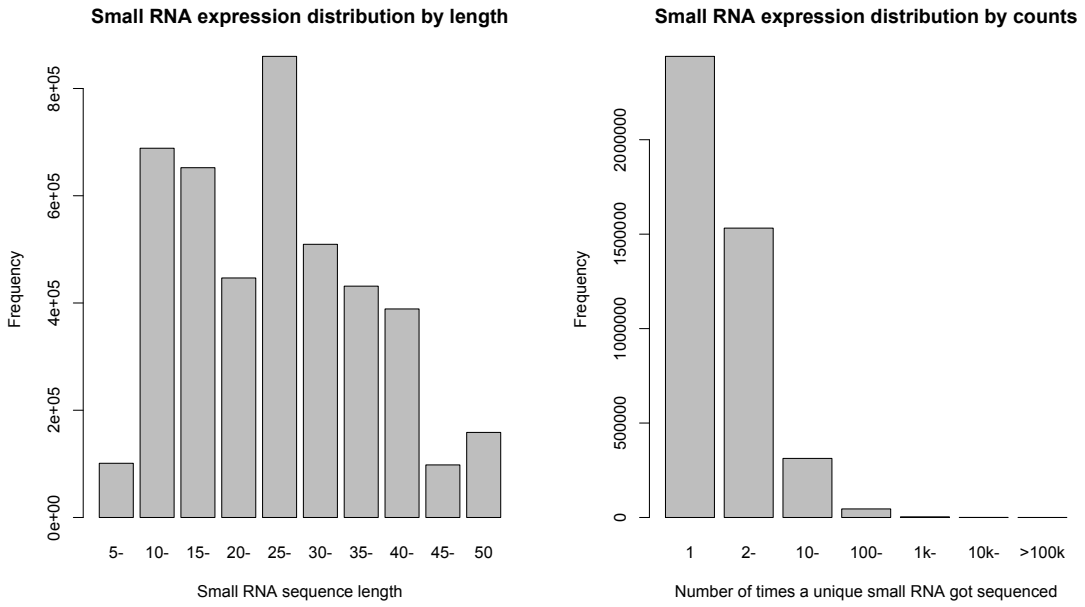


Figure 4.7.: Results of *X. bocki* small RNA sequencing show that the vast majority of small RNA transcripts have low sequencing counts (right).

from 2009 [Wheeler et al., 2009]. The small RNA dataset for *S. roscoffensis* consists of 2,740 different short sequences of length 17 to 25 nucleotides. The number of times a given sequence has been sequenced ranges from 1 to 678 (fig. 4.8).

4.5.3. MicroRNA detection in *Xenoturbella bocki*

The first step of the detection pipeline searches for mature miRNA candidates within the small RNA transcript data. I identified 6,621 sequences as potential candidates that could represent 29 bilaterian miRNA families (fig. 4.9, left). Each of these candidate sequences contains a stretch of 6 nucleotides within the first 10 bases that perfectly matches the seed sequence of the corresponding miRNA family. Different miRNA families show different levels of conservation. For each family I used a threshold representing the sequence similarity between member sequences I retrieved from miRBase. The candidate sequences needed to pass this minimum conservation threshold when compared to the reference sequence of their respective families. As expected, I found many more

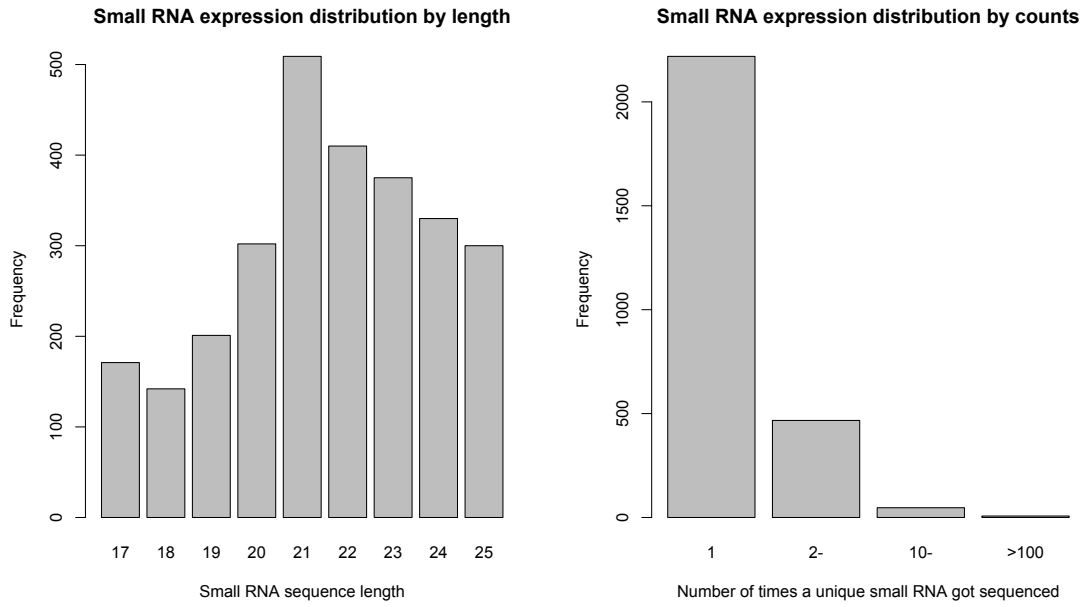


Figure 4.8.: Results of *S. roscoffensis* small RNA sequencing based on data from Wheeler et al. [2009] also show bias towards low sequencing counts of small RNA transcripts, but much reduced total counts (right) compared to our RNA sequencing of *X. bocki*.

candidates for families with lower conservation threshold. The only exception is *mir-252*, for which I found 600 candidates at a relatively high conservation threshold of 76.2% nucleotide identity. In comparison, *mir-31* only had 100 candidates requiring a threshold of 77.3%.

The second step matches the mature miRNA candidates to the genome and extracts the pre-miRNA candidates. I mapped each mature miRNA candidate to our draft genome of *X. bocki*. I had to reject mature miRNA candidates for 5 families as I was unable to find perfect matches within our genome. For the remaining 24 families I extracted the sequences surrounding the mature miRNA candidates as pre-miRNA candidates. The position of the mature miRNA was based on its position within the hairpin structure (5' or 3' end) with 10 nucleotides as buffer towards the closest end. The length of the pre-miRNA sequences was based on the maximum hairpin length of the mature miRNA's family according to the information extracted from miRBase. I found 9,620 pre-miRNA candidates to be tested for their folding structure (fig. 4.10, left). The number of mature

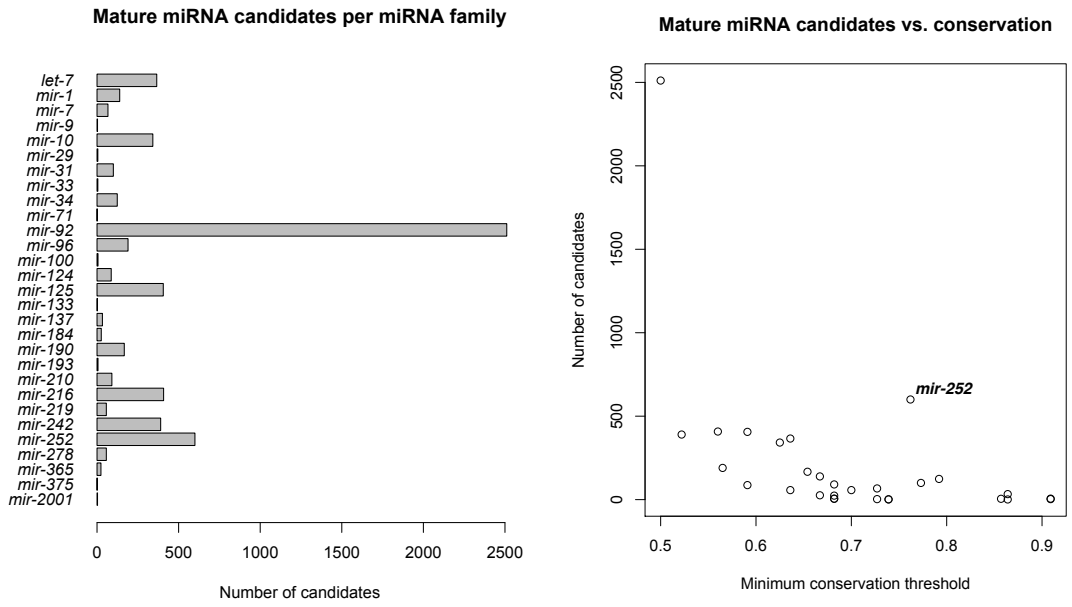


Figure 4.9.: Mature miRNA candidates identified from *X. bocki* small RNA data shows correlation between the conservation of a given family and the corresponding number of mature miRNA candidates, i.e. higher conservation rates likely result in fewer candidates (with *mir-252* as an outlier).

miRNA candidates does not correlate well with the number of extracted pre-miRNA candidates (fig. 4.10, right). This could be a result of a miRNA sequence occurring in several loci.

The last step evaluates the extracted pre-miRNA candidates according to their ability to form a viable hairpin structure. All 24 miRNA families for which I extracted pre-miRNA candidates have at least one sequences that is able to form a hairpin structure (fig. 4.11). Only miRNA families of *mir-137* and *mir-278* had no pre-miRNA candidates that achieved the highest grade of their folding structure (fig. 4.12).

For *X. bocki* I was able to confirm the existence of 17 miRNAs as reported by Philippe et al. [2011]. Only one family (*mir-278*) features a hairpin structure of less than ideal standard. Furthermore I was able to identify mature miRNA candidates for 4 miRNA

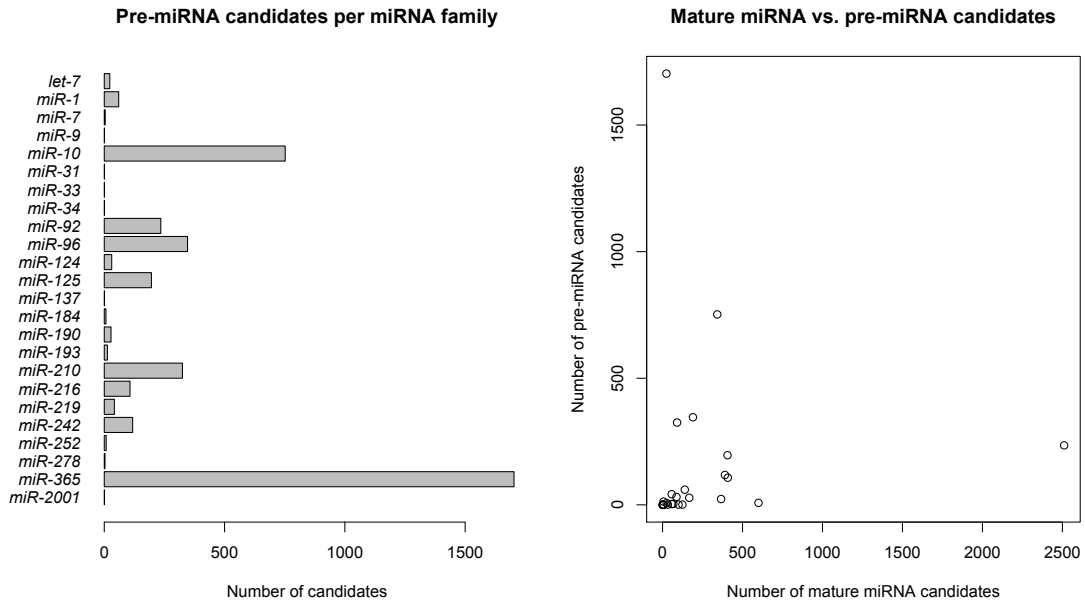


Figure 4.10.: Pre-miRNA candidates extracted from *X. bocki* genome based on mature miRNA candidates do not show a correlation between the number of mature miRNA candidates and the number of pre-miRNA sequences extracted from the genome, i.e. lower conservation thresholds and higher number of mature candidates do **not** correlate with an increased number of pre-miRNA candidates.

families that were only found within the genome before. Again, one of these families (*mir-137*) yielded a lower grade hairpin structure.

I was unable to confirm the existence of two reported families. For both, *mir-29* and *mir-100*, I was able to identify a mature miRNA candidate, but failed to match them to the genome.

4.5.4. MicroRNA detection in *Symsagittifera roscoffensis*

We have not yet sequenced small RNA data from our *Symsagittifera roscoffensis* specimens, but Kevin J. Peterson provided me with the small RNA data which was used in their previous analysis [Wheeler et al., 2009]. Of the 33 bilaterian miRNA families tested there they identified 7 families to be conserved within the acoel including 3 newly identified miRNAs compared to previous results in acoels [Sempere et al., 2007].

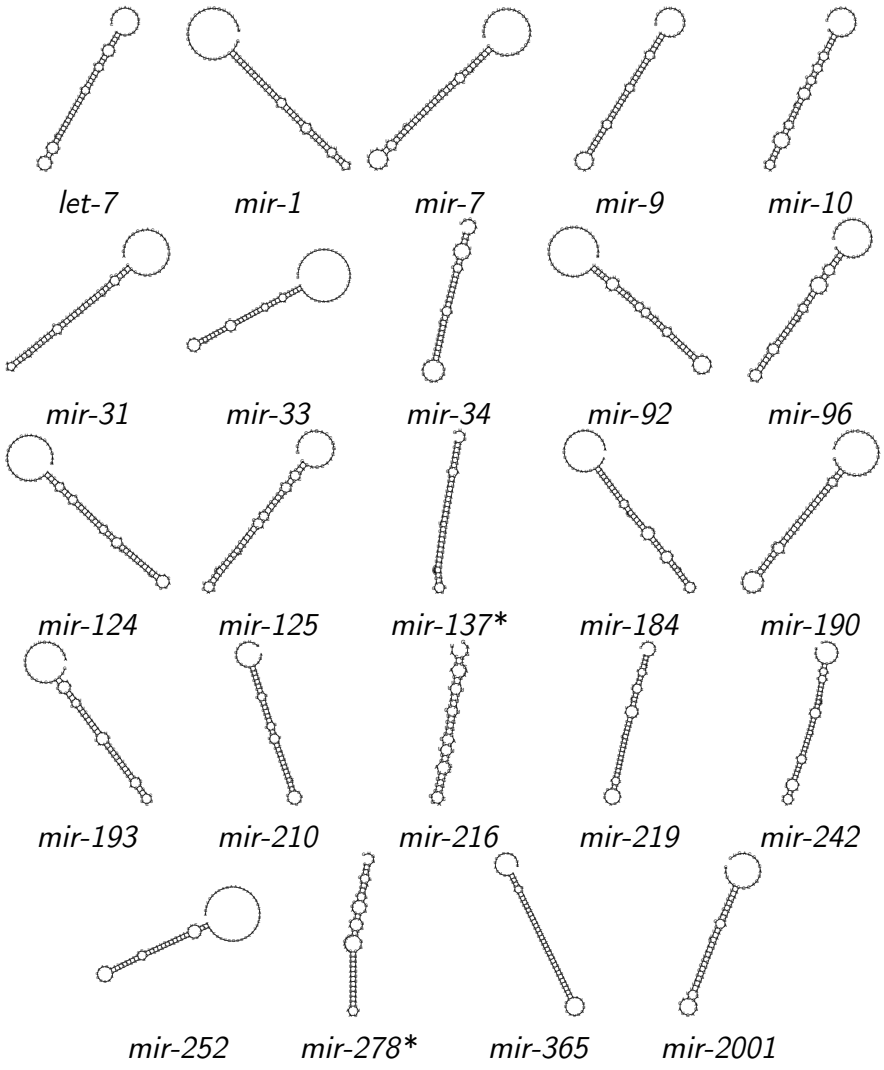
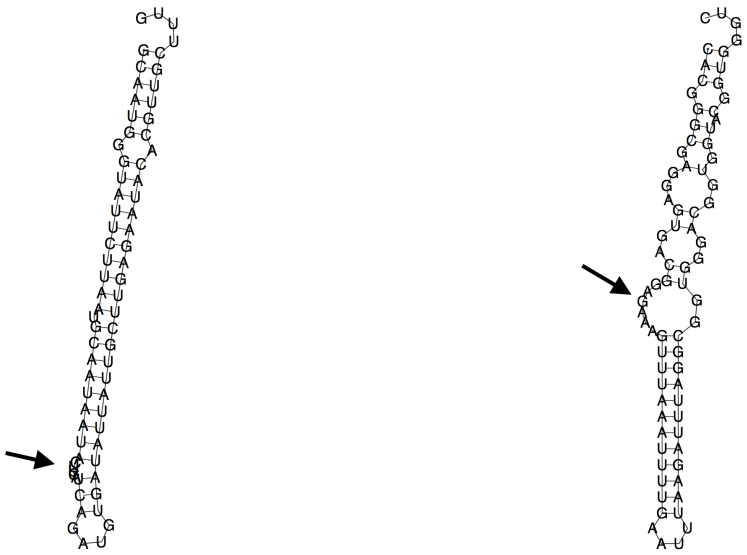


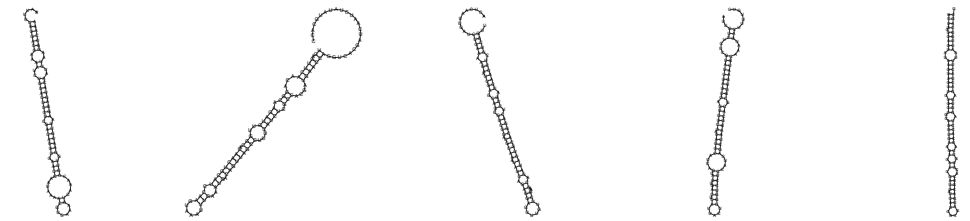
Figure 4.11.: Successful identification of viable bilaterian miRNA candidates in *X. bocki* from small RNA and genome data: secondary structures of best pre-miRNA candidates form viable hairpin structures that are able to be processed through the miRNA biogenesis pathway, * - hairpins with lower grading, i.e. hairpins more likely to result in lower Dicer efficiency.



mir-137 *mir-278*

Figure 4.12.: *X. bocki* miRNA detection results in detail: best pre-miRNA candidates from *X. bocki* that did not receive highest grading, arrows indicate bulges larger than the template used for evaluation (maximum of 3 consecutively unpaired nucleotides in the stem region).

I was unable to replicate all of the previous findings in *S. roscoffensis*. After running my pipeline on this dataset I was able to identify only 5 of the reported miRNAs with only 2 receiving the best hairpin grading (fig. 4.13).



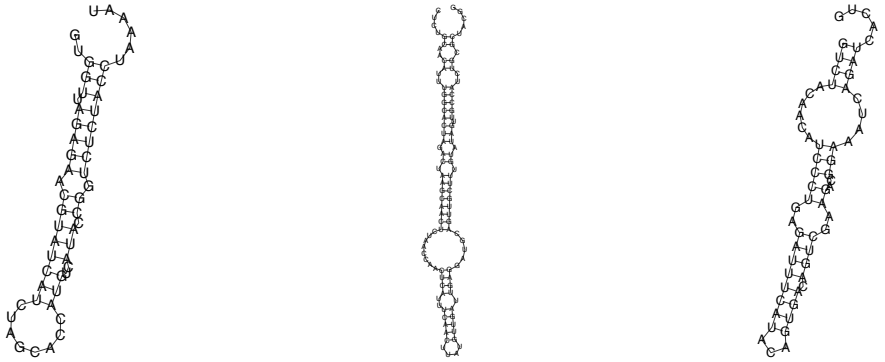
*mir-1** *mir-31** *mir-92* *mir-124** *mir-252*

Figure 4.13.: *S. roscoffensis* miRNA detection results based on small RNA data provided by Kevin J. Peterson and our genome data: best pre-miRNA candidates identified for 5 of the 7 families previously reported [Wheeler et al., 2009]. Detection of remaining families did not yield viable hairpins; * - lower grade hairpins.

Candidates for other reported families did not yield a positive result for various reas-

ons. For *mir-100* I was unable to find a mature candidate that passes the conservation threshold (90.9%). Unfortunately there are no data about the reported miRNA candidate. My assumption would be that their mature candidate falls below the threshold I established. For *mir-219* I was able to identify 2 mature candidates and 1 pre-miRNA candidate, but evaluation of the folding structure reported no viable hairpin formation.

I identified three additional miRNAs in the *S. roscoffensis* sequences (fig. 4.14). The pre-miRNA candidate for *mir-29* formed an ideal hairpin within our limits. Pre-miRNA candidates *mir-96* and *mir-125* were able to form hairpins, albeit of lower grading.



mir-29 *mir-96** *mir-125**

Figure 4.14.: Newly identified bilaterian miRNA families in *S. roscoffensis* based on small RNA and genome data: pre-miRNA candidates for *mir-96* and *mir-125* families contain bulge sizes greater than 3 unpaired nucleotides.

miRNA	2007[1]	2009[2]	2011[3]		new	
	Sro	*Sro*	*Hmi*	*Xbo*	*Xbo*	*Sro*[4]
let-7	-	-	-	X	X	-
mir-1	-	X	X	X	X	X*
mir-7	-	-	X	X	X	-
mir-9	-	-	-	X	X	-
mir-10	-	-	-	X	X	X
mir-31	X	X	X	X	X	X*
mir-33	-	-	-	X	X	-
mir-34	X	-	D	D	X	-
mir-92	X	X	X	X	X	X
mir-100	X	X	-	X	-	-
mir-124	X	X	X	D	X	X*
mir-125	-	-	-	X	X	X*
mir-133	-	-	-	-	-	-
mir-153	-	-	X	X	-	-
mir-184	-	-	-	X	X	-
mir-210	-	-	-	X	X	-
mir-219	X	X	X	X	X	-
mir-375	-	-	-	-	-	-
mir-252		X	X	X	X	X
mir-29			X	X	-	X
mir-96			X	D	X	X*
mir-137			-	D	X*	-
mir-190			X	X	X	-
mir-278			-	X	X*	-
mir-2001			X	X	X	-
mir-193					X	-
mir-216					X	-
mir-242					X	-
mir-365					X	-
mir-71					-	-
mir-78					-	-
mir-182					-	-
mir-183					-	-
mir-281					-	-
mir-315					-	-
mir-981					-	-

Table 4.2.: Presence of bilaterian miRNA families in Xenacoelomorpha, X - detected, X* - detected with lower grading, "-" - not detected, D - genomic traces (but absent from small RNAs), empty - not (indicated as) tested; *Sro* - *Symsagittifera roscoffensis* (aceol), *Hmi* - *Hofstenia miamia* (acoel), *Xbo* - *Xenoturbella bocki* (xenoturbellid); data from: [1] - Sempere et al. [2007], [2] - Wheeler et al. [2009], [3] - Philippe et al. [2011], new - miRNA identification from small RNAs presented in this thesis, [4] - RNA data from Wheeler et al. [2009]

4.6. Discussion

Research on miRNAs, on what differentiates nucleotide sequences that act as miRNAs from those that do not and how to use these information for accurate detection is still in its infancy. Throughout my research on this topic, I noticed a focus on detecting the presence of miRNAs, but a lack of experimental evidence about the function of the found sequences. Advances in sequencing technology and the easy application of miRNA detection methods have made it possible to investigate the presence of miRNAs or miRNA-like sequences in a broad range of species. I found that only few of these publications feature experiments to corroborate the function of preserved sequences or at least apply target prediction methods to assess the potential role of the found miRNAs. It is understandable, that additional perturbation experiments are costly, but, unfortunately, miRNA target predictions alone might not be reliable in assessing function due to a high potential for false positive results [Pinzón et al., 2017]. Without more information about the role of miRNAs across the Metazoa or Bilateria their implied importance is based solely on their preservation. A notable exception to this is the research on miRNAs which have been linked to human diseases (review in Paul et al. [2017]).

MiRNA detection has yet to reach its full potential. MiRscan [Lim et al., 2003] and miRDeep2 [Friedländer et al., 2012] are some of the most commonly used miRNA identification tools with over 1500 and 900 citations, respectively. Despite the efforts that went into developing these methods, they are both unable to confirm all miRNAs that had been sequenced before the methods' publications. MiRscan was only able to recover 50% of the known *C. elegans* miRNAs, while miRDeep2's sensitivity depended on the species ranging from 71% (sea squirt) to 90% (sea anemone). If we assume these results to be representative for other genomes, we have to concede the fact, that we expect a false negative rate of at least 10-50%. Similarly, my own approach failed to detect 2 of the 19 previously sequenced bilaterian miRNAs in *X. bocki* and 1 of the 6 bilaterian miRNAs found in *S. roscoffensis*. In the following chapter, I will show how some of my negative results can be explained by my choice of detection parameters in an effort to reduce false positives. These shortcomings highlight the potential for improvement through continued research on miRNAs.

The advancement of RNA folding methods may call previous results into question. I searched for useful miRNA characteristics not just by looking at other miRNA identific-

ation methods, but also at the published sequences itself. MiRBase contains sequenced pre-miRNAs that are useful to determine the structure of RNA hairpin foldings with a focus on miRNA detection. Given the same sequence, I have noticed that many of the proposed hairpin structures on miRBase are different from hairpin structures computed by RNAfold [Lorenz et al., 2011] (fig. 4.15). This also implies that methods that have used previous versions of RNA folding software to train their identification pipeline could perform differently now. More research is needed to ensure which miRNA characters are not only useful, but can also be reproduced invariably.

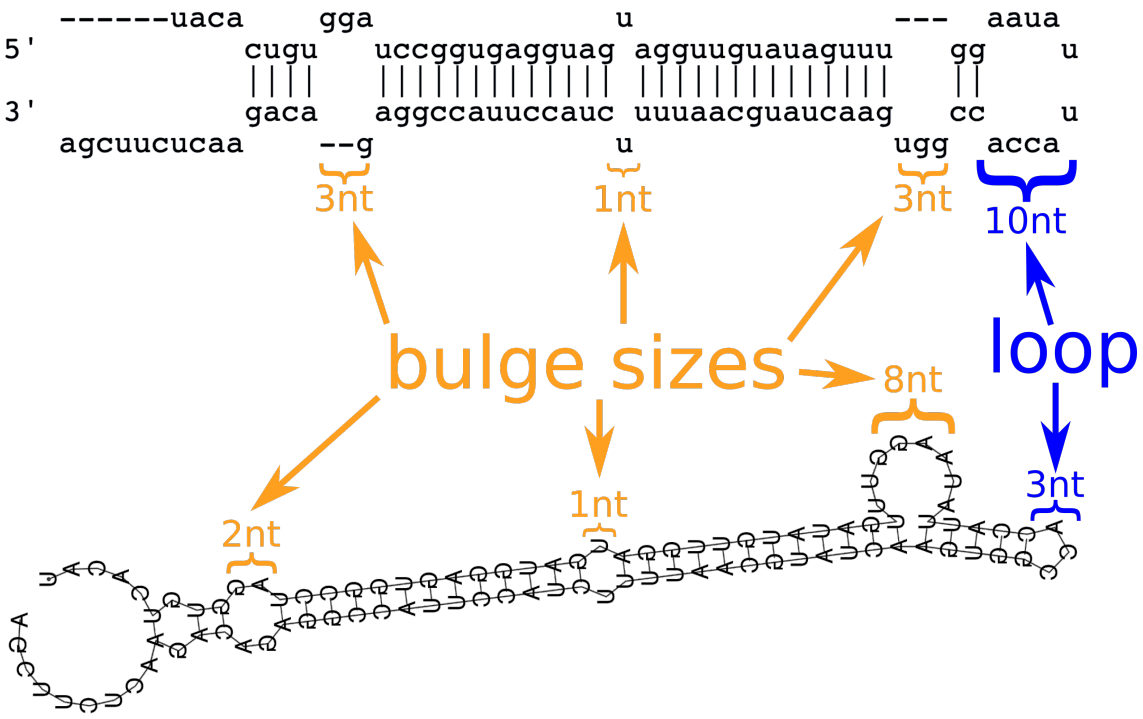


Figure 4.15.: Comparison of RNA folding structures computed by different methods. Nucleotide sequences represent the same pre-miRNA of *let-7* found in *C. elegans*, but hairpin structures show different bulge and loop sizes between miRBase display (top, folding algorithm not listed) and a folding I computed (bottom) using RNAfold (version 2.4.3, default parameters, Lorenz et al. [2011]).

5. Prediction of microRNA candidates from xenacoelomorph genomes

5.1. Motivation

Presence and absence of miRNAs in Xenacoelomorpha has been used to support their previously established phylogenetic positions. Early studies sequenced miRNAs in acoels and compared these to other bilaterians. Sempere et al. [2007] identified 16 miRNA families conserved between Protostomia and Deuterostomia. They sequenced only 6 out of these 16 in acoel species which supported the notion of Acoela positioned as sister to all remaining Bilateria. Wheeler et al. [2009] increased the bilaterian miRNA set to 34, but only sequenced an additional 2 miRNAs in acoels (1 of which was previously untested). They followed the interpretation of most bilaterian miRNAs evolving after the split from acoels.

Additional investigations have revealed a larger bilaterian miRNA complement in Xenacoelomorpha. Philippe et al. [2011] sequenced small RNA from the acoel *Hofstenia miamia* and found 10 bilaterian miRNAs previously undetected in acoels. They also sequenced the xenoturbellid *Xenoturbella bocki* (sister to acoels and nemertodermatids) and found not only the aforementioned 10 miRNAs, but added 8 more bilaterian miRNAs to the total bilaterian complement of Xenacoelomorpha.

A broader species sampling as well as improved sequencing techniques and identification methods continue to close the gap between conserved miRNAs identified in Xenacoelomropha and other Bilateria. I was able to show evidence for even more con-

servation using our new *X. bocki* small RNA data and my identification pipeline (previous chapter, table 4.2).

We believe that the absence of bilaterian miRNAs from xenacoelomorphs could be artifactual. The small taxon sampling, especially within Xenoturbellida, and the lack of RNA transcripts across Xenacoelomorpha prevents us from accurately estimating the total miRNA complement. We are currently missing RNA data from most xenacoelmorph species and we only have one set of RNA data from *X. bocki*. This problem is exacerbated by the fact, that miRNAs can be involved in development and their expression can fluctuate [Wienholds et al., 2005]. A single snapshot from the adult stage of *X. bocki*, that we currently have, may not capture all miRNAs prevalent within our species of interest.

Small RNA results can be complemented by a search for miRNA candidates within the genome. Most miRNA identification methods require small RNA data to identify viable miRNAs and their associated hairpin structures. In their analysis, Philippe et al. [2011] reported 4 miRNAs for which they did not sequence RNA, but identified potential candidates from the genome. I was able to confirm the presence of all 4 of these miRNAs using our new small RNA dataset, proving the feasibility of miRNA predictions based on genomic data alone.

Successful prediction of miRNA candidates from the genome would be very useful for hard to sample species such as ours. The lack of readily available specimens and the issues experienced during sequence extraction on small and rare specimens makes the examination of xenacoelomorphs challenging. A miRNA pipeline that relies solely on genomic data would decrease the amount of time and resources spent. Additionally we would be able to process already available genomes without the need for new sequencing.

The number of potential mature miRNA candidates necessitates an efficient computational approach. The initial criterion to identify a mature miRNA candidate is the seed sequence which consists of only 6 nucleotides. The short length of the sequence means, it can be quite ubiquitous within the whole genome of an animal purely by chance. I created a script that enables me to scan the genome of my species of interest for perfect matches to miRNA seed sequences. It will then extend the sequence to full mature sequence length and keep them, if the candidate matches the miRNA family's reference sequence at a given threshold (expressed as nucleotide identity).

Mature miRNA candidates predicted from the genome must adhere to miRNA characteristics. I subject the candidates extracted from the genome to the same treatment as candidates from small RNA data. This will reject candidates that would not be able to comply with miRNA processing. The remaining candidates would then, if transcribed, be able to form a hairpin structure from which a functional mature miRNA can be released.

However, we cannot rely on the existence of potential miRNA candidates without ascertaining how reliable these predictions are. I devised a set of negative controls using both generated and real miRNA data to establish the failure rate of this prediction approach. I use our draft genome of *X. bocki* as a study case. I added the published genome of *Nematostella vectensis* as a second data set to see how these tests perform on a larger genome.

The results of these test show that my pipeline effectively reduces the number of false positive predictions. This increases my confidence in the validity of the bilaterian miRNAs I predict in our xenacoelomorph genomes.

5.2. Prediction pipeline

The aim of my prediction pipeline is to identify genomic sequences to complement my results from RNA data. Genomic sequences that are similar to mature miRNA sequences in other species represent potential candidates, if these sequences where transcribed. However, akin to sequences identified in small RNA data, we cannot rely purely on sequence similarity alone. The genomic candidate must be scrutinised for its compliance with the miRNA biogenesis pathway. The requirements I set for successfully identifying a genomic candidate are the same as previously established for candidates from small RNA:

- The mature miRNA candidate must contain the perfectly conserved seed sequence to be considered member of a given miRNA family.
- The mature miRNA candidate must be similar to the corresponding miRNA family's reference sequence. This similarity must pass a set threshold.

- The mature miRNA candidate must be embedded in a larger sequence, the pre-miRNA, which must be able to form a hairpin structure to be a viable candidate for miRNA biogenesis.

The difference between the identification pipeline and the prediction pipeline is the initial source for mature miRNA candidates. For the identification pipeline I use small RNA transcripts which are then mapped to the genome. For the prediction pipeline I use the genome directly to search for potential miRNA candidates in the absence of small RNA transcript data.

The previously described identification pipeline follows these steps:

1. Find mature miRNA candidate among small RNA transcripts
2. Map mature miRNA candidate to genome
3. Extract pre-miRNA candidate sequence containing mature miRNA candidate
4. Compute pre-miRNA candidate folding structure
5. Evaluate hairpin structure formed by pre-miRNA candidate

For the prediction pipeline I substitute steps 1 and 2 with a direct scan of the genome using a custom Python script (implementation: `microRNA_prediction_from_DNA_kmers.py`, see Appendix B.2). My pipeline allows for skipping steps, if the resulting data has already been computed. For the prediction pipeline I created a script that mimics the extraction of mature miRNA and pre-miRNA candidates using the genome instead of small RNA transcripts (fig. 5.1).

The first step of the prediction pipeline searches for seed sequences in the genome. All potential mature miRNA candidates must include the perfectly conserved seed sequence of the family of interest. I created a custom Python script which uses text matching to scan each sequence in an assembled genome for all instances of the seed sequence.

The second step extends the seed sequence to adjacent nucleotides that would encompass the potential mature miRNA candidate and compares this longer sequence to the family's reference sequence. I extend the seed sequence according to the family's mature miRNA length which I gathered from the family's member sequences (as listed in miRBase). Originally my pipeline only kept predicted mature miRNA candidates, if they

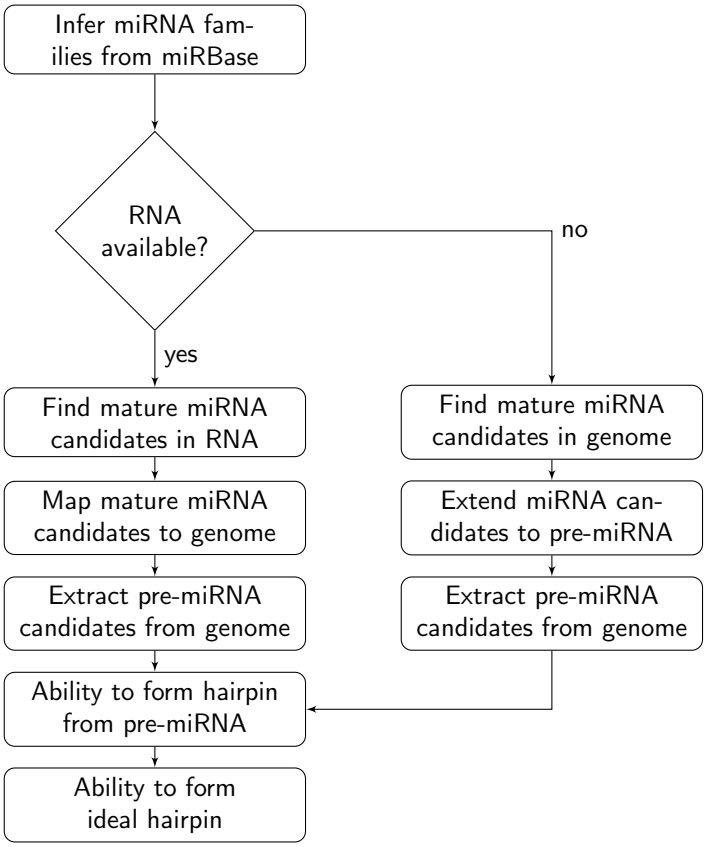


Figure 5.1.: MiRNA identification (left) steps to identify and validate miRNA candidates. MiRNA candidate prediction (right) reuses hairpin evaluation steps to validate candidates from genome.

pass the respective family's minimal threshold for nucleotide identity when compared to the family's reference sequence. This was in accordance with my restrictions I used for identifying candidates from small RNA transcripts. I later discovered that this led to the exclusion of previously found miRNAs as well as an increased potential for predicting false positives if the family threshold is too low. I address these findings in more detail in the results and the negative controls section of this chapter.

The last step of the prediction pipeline extends the mature miRNA candidates to pre-miRNA candidates. The mature miRNA sequence is extended based on the respective miRNA family's hairpin size and the acting strand (again inferred from miRBase data) including a stretch of 10 nucleotides between the mature miRNA candidate and the respective end of the pre-miRNA sequence. This pre-miRNA candidate is then stored alongside the mature miRNA candidates.

The resulting pre-miRNA candidate is now folded using RNAfold (version 2.4.3, default parameters, Lorenz et al. [2011]) and evaluated reusing my miRNA identification pipeline. The pre-miRNA candidate must be able to form a viable hairpin in order to allow for processing according to the miRNA biogenesis pathway. I evaluate the hairpin structure reusing the hairpin evaluation module of my identification pipeline, i.e. grading according to correct number of stem regions, correct placement of mature miRNA candidate and maximum number of consecutively unpaired nucleotides (bulges and loop sizes).

I encountered computational problems when predicting mature miRNA and pre-miRNA candidates from genomes with longer assembled sequences. I experienced a slowdown of the mature miRNA prediction step when applied to the cnidarian genome of *N. vectensis*. I identified the problem to be the longer sequences in the assembled genome of *N. vectensis* (N50 $>$ 472kb) compared to the *X. bocki* genome (N50 $\approx$ 62Kb). I noticed that the execution of individual Python string operations, such as searching for or replacing substrings and reversing sequences, increases exponentially with the length of the sequence used (fig. 5.2). This becomes especially noticeable in sequences that contain 1 million bases or more. In our current draft genome of *X. bocki* there is only one sequence that is close to this length (960,978 nucleotides) while there are 66 sequences in the assembled genome of *N. vectensis* which are longer (longest: 3,256,212 nucleotides). This would make the application of my pipeline unfeasible for genome assemblies of higher quality.

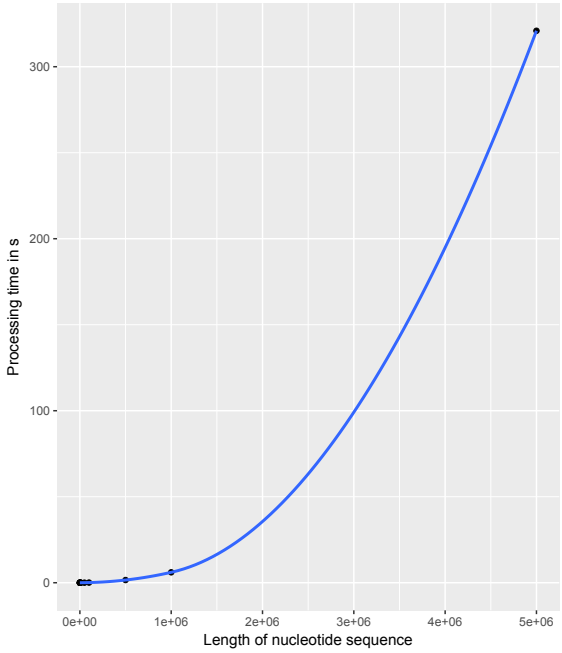


Figure 5.2.: Performance loss in longer sequences. An increase in sequence length (x-axis) exponentially increases the processing time (y-axis) of executed Python string operations.

I remedied the increase in string operation execution time by creating smaller subsequences of each genome sequence. I extracted all partial sequences of 6 nucleotides (6-mers, using a sliding window) keeping information about each fragment's origin (position and genome sequence name, implementation: `split_genome_into_miRNA_kmers.py`, see Appendix B.2). Then I filtered these fragments for those that match any of the family seed sequences I am interested in. From these I gathered the information about the potential mature miRNA candidate's position within each genome and continued to extend and compare the sequence as described above. This allowed me to shorten the total runtime (prediction and hairpin evaluation) to well under a day for each genome and miRNA families dataset combination.

High conservation of mature miRNAs makes prediction feasible, but false positive rates have to be estimated. Mature miRNAs are highly conserved, presumably due to their importance in gene regulation and the potentially drastic effects of small changes on mRNA target identification. I added the evaluation of pre-miRNA hairpin structures to increase the validity of the inferred mature miRNA. However, it is not clear if my

predictions are accurate or if they could be a result of random chance. I devised a suite of negative controls to test the probability of predicting false positives (detailed in the corresponding section of this chapter).

5.3. Predictions of bilaterian microRNA candidates in xenacoelomorphs

The first miRNA studies in xenacoelomorphs did not find evidence for most of the miRNAs conserved between protostomes and deuterostomes [Sempere et al., 2006, 2007, Wheeler et al., 2009]. Only 8 of the more than 30 miRNAs tested were sequenced in studies of 2 acoel species. This apparent lack of bilaterian miRNAs was used to support the position of Acoela as sister to all remaining bilaterians. The larger miRNA complement common to both Protostomia and Deuterostomia was thought to have evolved after the split from the acoels. If this lack of bilaterian miRNAs is correct the inferred position of Xenacoelomorpha as part of the Deuterostomia [Philippe et al., 2011] would imply a substantial loss of the bilaterian miRNA complement.

The loss of conserved miRNAs is considered a rare evolutionary event. Previous studies have shown a high conservation of miRNA families within animals [Hertel et al., 2006, Sempere et al., 2006, Prochnik et al., 2007]. Many mature miRNA sequences were also highly conserved even between distantly related species. The presence and absence of tested miRNAs aligns with the overall species phylogeny. These aspects imply that miRNAs have accumulated over time with new miRNAs marking major divergences in evolution. The high conservation and implied importance of miRNAs has led to the interpretation of miRNA losses to be rare evolutionary events.

Continued research has already shown the existence of some of the previously missing miRNAs in xenacoelomorphs. Philippe et al. [2011] sequenced 10 bilaterian miRNAs in the acoel *H. miamia* and the xenoturbellid *X. bocki* that were previously stated as absent from acoel species. 8 more bilaterian miRNA sequences were found in *X. bocki*. I confirmed the existence of all of these miRNAs using my newly developed method and new sequencing data from *X. bocki* (see previous chapter). I successfully identified and validated RNA candidates for 4 bilaterian miRNA families that were previously only predicted from the *X. bocki* genome. Furthermore, I was able to provide evidence for 4

additional bilaterian miRNAs in *X. bocki* that have not been sequenced or tested before (table 4.2). Using *S. roscoffensis* RNA data from Wheeler et al. [2009] I identified and validated miRNA candidates for 9 of the 36 tested bilaterian miRNA families.

The continued absence of bilaterian miRNAs might be an artefact of the difficulty acquiring xenacoelomorph data. Xenacoelomorphs are generally hard to sample, which is one reason why we currently only have a single high quality dataset of sequenced *X. bocki* small RNAs. We lack data from different developmental stages from *Xenoturbella* as well as good quality acoel small RNAs. The low levels of expression of some miRNAs (e.g. due to temporal or spatial specificity) makes it even more difficult for us to capture a complete picture of the miRNA toolkit. These fragmentary data could cause an absence of evidence of miRNA sequences which would be interpreted as absence from the organism or clade. The prediction of miRNAs from genomic data allows us to add complementary miRNA information for which we do not yet have RNA sequences.

I have shown that genomic predictions of miRNAs can be successful. Philippe et al. [2011] found genomic traces of 4 bilaterian miRNAs in the genomes of *H. miamia* and *X. bocki*, but no RNA evidence. I was able to confirm the existence of these miRNAs in *X. bocki* using our new RNA dataset (see previous chapter). However, I failed to detect 3 miRNAs that have been previously sequenced (table 4.2). For these sequences I should be able to find genomic evidence using my prediction pipeline.

5.3.1. Prediction results from *Xenoturbella bocki*

I was unable to detect miRNA candidates for 12 bilaterian miRNA families using our newly sequenced *X. bocki* RNA data (table 4.2). My detection pipeline was able to predict mature candidates for all of these families after finding a perfect match of their respective seed sequence in our *X. bocki* genome.

All bilaterian miRNA families returned at least one viable hairpin structure. Predicted mature miRNA candidates from the genome were extended to pre-miRNA length, folded and evaluated. I rejected all candidates that are unable to form a viable hairpin according to the known requirements of the miRNA biogenesis pathway. For each miRNA family, I predicted at least one mature candidate that was also part of a viable hairpin structure (fig. 5.3).

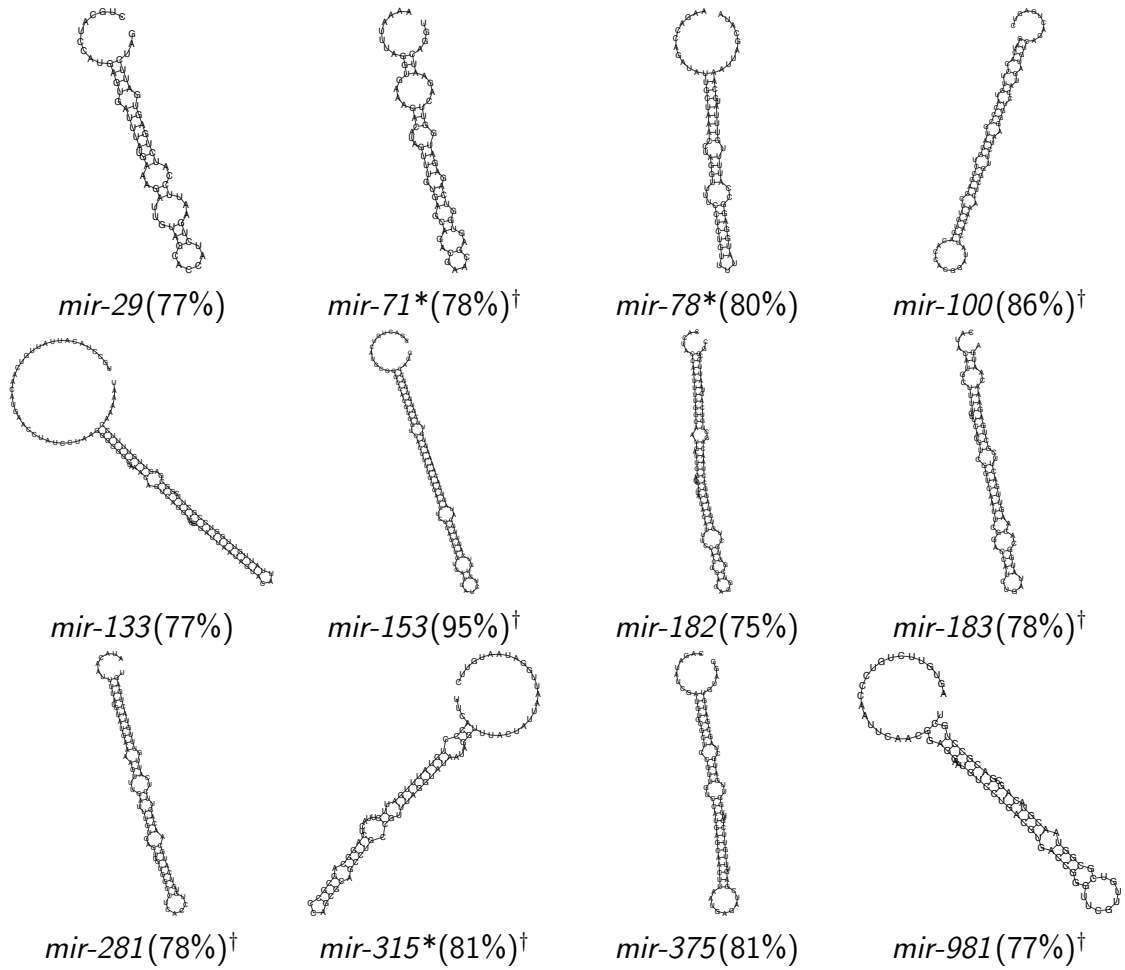


Figure 5.3.: Best pre-miRNA candidates predicted from *X. bocki* genome for bilaterian miRNA families, percentages in parentheses display conservation between the mature miRNA candidate and the family's reference sequence; * - lower grade hairpin, † - best mature candidate below family minimum conservation threshold.

Conservation of best mature candidates varied between families. When I compare all the mature miRNA candidates to their respective family's reference sequence I found between 75% and 95% nucleotide identity. I investigated each family and the predicted miRNA candidates closer and found differences in reliability.

mir-29 has previously been sequenced in *H. miamia* and *X. bocki* [Philippe et al., 2011]. I was able to identify a viable hairpin, which contains a mature miRNA candidate of 77% conservation compared to the human *mir-29a* sequence. This candidate does

not match the previously identified *mir-29* sequence [Philippe et al., 2011]. I computed the conservation rate for the reported sequence and found it to be equal to my best candidate. I did not find the previously reported sequence in my set of mature candidates. Furthermore, using a text search (grep), I scanned both my small RNA sequences and the genome of *X. bocki*, but I was unable to find the reported sequence (or its reverse complement). The absence of the sequence from the genome implies that the previously sequenced miRNA candidate was a contamination. The similarity between my predicted miRNA candidate and the *mir-29* family passes the threshold and the corresponding pre-miRNA sequence is able to form a viable hairpin.

mir-100 has been sequenced several times in different xenacoelomorph species [Sempere et al., 2007, Wheeler et al., 2009, Philippe et al., 2011]. From the *X. bocki* genome I predicted a viable hairpin. The corresponding mature candidate shows high conservation: 86% nucleotide identity compared to the human *mir-100* sequence. The previously reported *X. bocki* sequence [Philippe et al., 2011] has a lower conservation rate of 82%. I found the reported sequence within my small RNA data, but, akin to *mir-29*, I did not find a match of the reported sequence or its reverse complement within the *X. bocki* genome. However, I did find my best mature candidate expressed within the small RNA data. The exclusion of both the reported and the newly sequenced best candidate from a successful identification was due both of their conservation being lower than the conservation threshold of 90.9% for the *mir-100* family. This shows that my adhering to a family threshold can be too conservative removing viable candidates from consideration. The use of my miRNA prediction pipeline can provide candidates which can then be confirmed through the available RNA data.

mir-153 has been reported in xenacoelomorphs [Sempere et al., 2007, Wheeler et al., 2009, Philippe et al., 2011]. I predicted a viable hairpin including a mature miRNA with a very high conservation of 95%. Unlike the above examples, I was able to identify the exact same sequence from the *X. bocki* genome that had previously been sequenced from RNA [Philippe et al., 2011]. This sequence also appears in the small RNA sequence data. The reason my identification pipeline failed in this instance was that the mature candidate with 95.3% conservation fell just below the conservation threshold I identified for the *mir-153* family: 95.5%. Together with the previous example (*mir-100*) this shows that thresholds can be set too conservatively when using only information provided by sequences from miRBase. This causes an exclusion of truly present miRNA sequences. A

more lenient threshold would amend this, but potentially introduce more false positives. It is also unclear, as to how such a threshold could be chosen, if it is not based on previously sequenced and confirmed sequences.

MiRNA families *mir-133* and *mir-375* have been previously tested, but no mature candidates have been sequenced [Sempere et al., 2006, 2007, Wheeler et al., 2009, Philippe et al., 2011]. I predict viable hairpins for both families. I found several mature candidates for *mir-133* that achieve a conservation level of 77% (higher than *mir-133*'s 73.9% conservation threshold), but none of them are present in the small RNAs. *mir-375* predictions contain mature candidates that are also above the family's conservation threshold (81% compared with 73.9%). However, there is also no evidence for them in the small RNA data.

The hairpins and the included mature miRNA candidates for the previously untested families either fall below their families' thresholds (*mir-71*, *mir-183*, *mir-281*, *mir-981*), form hairpins of lower grade (*mir-78*) or both (*mir-315*). The only exception is the best candidate for *mir-182*. The mature candidates for *mir-375* are the only ones with a conservation of more than 80%, but without any evidence from the small RNA. The low conservation rate or lower hairpin grading of the best candidates increases the chances for these predictions to be false positives.

5.3.2. Predictions from acoel genomes

For *Symsagittifera roscoffensis* I applied my miRNA prediction pipeline on the 27 families for which I was unable to identify a viable miRNA candidate from the small RNA transcripts (table 4.2). I predict viable hairpin structures for all tested families (fig. 5.4). 9 mature miRNA candidates are below their family's minimum conservation. 11 of the best candidates have a conservation of 80% or more.

For *Paratomella rubra* I ran the miRNA prediction pipeline on our draft genome for all 36 bilaterian miRNA families. We currently do not possess any kind of small RNA data from this acoel. Results from a future sequencing effort could be compared to the predictions made here in order to validate my findings. Akin to *S. roscoffensis* I was also able to predict viable hairpin structures for every miRNA family tested (fig. 5.5). 19 families had mature candidates with a conservation of 80% or more, but 12 candidates failed to reach their family's minimum conservation threshold.

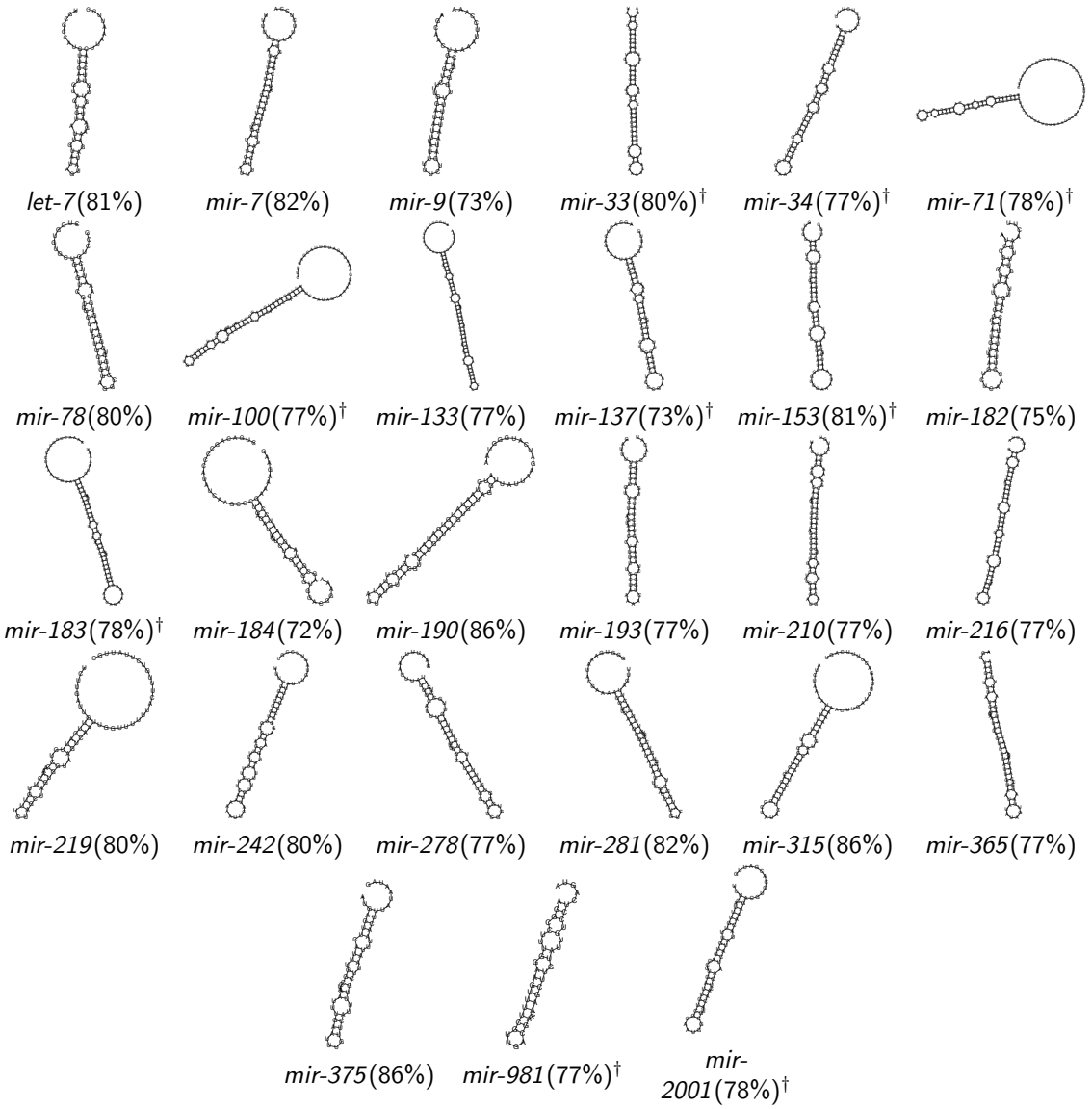


Figure 5.4.: Best pre-miRNA candidates predicted from *S. roscoffensis* genome for bilaterian miRNA families, percentages in parentheses display conservation between the best mature miRNA candidate and the family's reference sequence; * - lower grade hairpin, † - best mature candidate below family minimum conservation threshold.

Several of the predicted pre-miRNAs contain mature miRNA sequences of exceptionally high conservation (90-95% nucleotide identity). *mir-7*, *mir-34*, *mir-124* and *mir-219* have all been sequenced and identified several times, including using my own pipeline. This demonstrates the capabilities of my pipeline to predict candidates from genomic data.

Comparing the miRNA predictions from both acoel species reveals putative miRNA candidates of high conservation. *let-7* is a miRNA highly conserved between protostomes and deuterostomes, but was noticeably absent in studies of acoels [Sempere et al., 2006, 2007]. This was interpreted as a key miRNA missing from Acoela to support the sister relationship to all other bilaterians. *let-7* was eventually sequenced in *X. bocki* [Philippe et al., 2011]. I was also able to predict *let-7* miRNA candidates in both *P. rubra* and *S. roscoffensis*, which further refutes the notion of Xenacoelomorpha lacking *let-7*.

MiRNA predictions match reported presence of miRNAs in Acoela. 14 bilaterian families have been sequenced from acoel species [Sempere et al., 2006, 2007, Wheeler et al., 2009, Philippe et al., 2011]. I was able to predict viable mature miRNAs and corresponding pre-miRNAs for both *S. roscoffensis* and *P. rubra*, from which I do not have high quality small RNA data. With the exception of one family (*mir-2001*), all predicted candidates surpass the minimum conservation of their respective family (table 5.1).

Predictions from across Xenacoelomorpha increase my confidence in miRNA presence. Only two previously tested miRNA families (*mir-133* and *mir-375*) have been consistently absent from all investigated xenacoelomorph species. I was also unable to find good mature candidates from the *X. bocki* RNA data (previous chapter). But my predictions show viable candidates, that almost all surpass the families' minimum conservation in all xenacoelomorph genomes tested in this study (table 5.1). This increases my confidence that the absence of these miRNAs from RNA data could be explained by a low or temporal expression, that we were unable to capture.

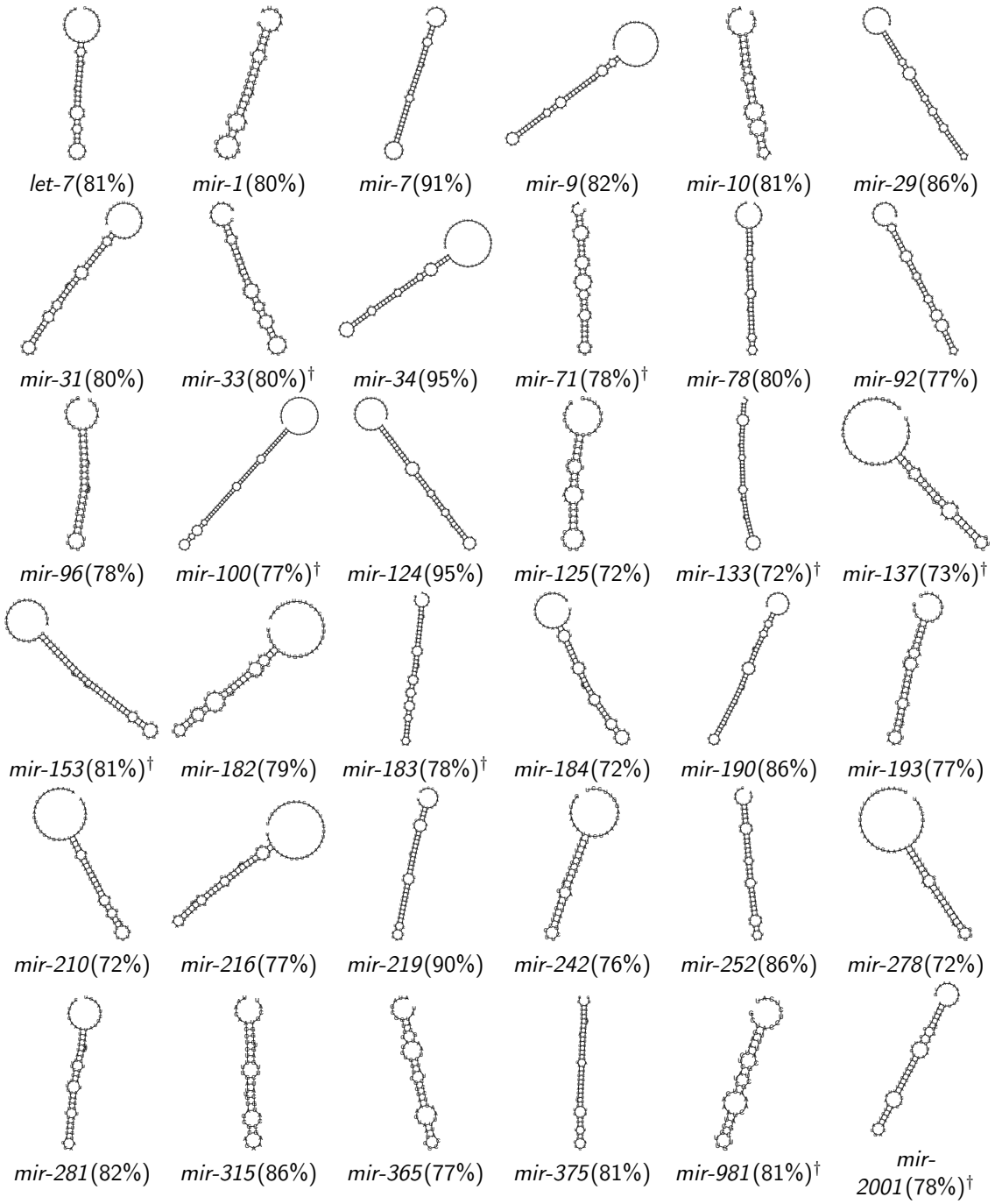


Figure 5.5.: Best pre-miRNA candidates predicted from *P. rubra* genome for bilaterian miRNA families, percentages in parentheses display conservation between the best mature miRNA candidate and the family's reference sequence; * - lower grade hairpin, † - best mature candidate below family minimum conservation threshold.

miRNA	2007[1]	2009[2]	2011[3]		new/predicted		
	Sro	*Sro*	*Hmi*	*Xbo*	*Xbo*	*Sro*[4]	*Pru*
let-7	-	-	-	X	X	P	P
mir-1	-	X	X	X	X	X*	P
mir-7	-	-	X	X	X	P	P
mir-9	-	-	-	X	X	P	P
mir-10	-	-	-	X	X	X	P
mir-31	X	X	X	X	X	X*	P
mir-33	-	-	-	X	X	P†	P†
mir-34	X	-	D	D	X	P†	P
mir-92	X	X	X	X	X	X	P
mir-100	X	X	-	X	X†	P	P†
mir-124	X	X	X	D	X	X*	P
mir-125	-	-	-	X	X	X*	P
mir-133	-	-	-	-	P	P	P†
mir-153	-	-	X	X	X†	P	P†
mir-184	-	-	-	X	X	P	P
mir-210	-	-	-	X	X	P	P
mir-219	X	X	X	X	X	P	P
mir-375	-	-	-	-	P	P	P
mir-252		X	X	X	X	X	P
mir-29			X	X	P	X	P
mir-96			X	D	X	X*	P
mir-137			-	D	X*	P†	P†
mir-190			X	X	X	P	P
mir-278			-	X	X*	P	P
mir-2001			X	X	X	P†	P†
mir-193					X	P	P
mir-216					X	P	P
mir-242					X	P	P
mir-365					X	P	P
mir-71					P†	P†	P†
mir-78					P	P	P†
mir-182					P	P	P
mir-183					P†	P†	P†
mir-281					P†	P	P†
mir-315					P†	P	P†
mir-981					P†	P†	P†

Table 5.1.: Presence of bilaterian miRNA families in Xenacoelomorpha including predictions, X - detected, † - conservation of mature candidate below family conservation minimum, "-" - not detected, D - genomic traces (but absent from small RNAs), P - predicted from genome, empty - not (indicated as) tested, * - best hairpin does not have perfect grading; *Sro* - *Symsagittifera roscoffensis* (aceol), *Hmi* - *Hofstenia miamia* (acoel), *Xbo* - *Xenoturbella bocki* (xenoturbellid), *Pru* - *Paratomella rubra* (acoel); [1] - Sempere et al. [2007], [2] - Wheeler et al. [2009], [3] - Philippe et al. [2011], new/predicted - miRNA identification & prediction presented in this thesis, [4] - RNA data from Wheeler et al. [2009]

5.4. Negative controls of the prediction pipeline

Many prediction results within the Xenacoelomorpha are of low conservation, i.e. the best mature miRNA candidate for a bilaterian miRNA family has less than 80% nucleotide identity compared with the family's reference sequence. Data from miRBase shows that conservation can vary between species and between miRNA families (table 4.1). However, low conservation thresholds increase the chances of finding similar sequences by chance.

I therefore devised a series of negative controls to test my prediction pipeline. My approach to test the false positive rate for my predictions is to estimate the number of positive results we get from miRNA sequences that are not expected to be present within the Xenacoelomorpha. I use the genome of *X. bocki* (genome size: 120Mb) for these negative controls. Additionally, I also tested the prediction accuracy on the genome of the cnidarian *Nematostella vectensis* (genome size: 356Mb). This allows me to compare the test results for different genome sizes as a given random sequence is more likely to be found in a larger genome (false positive). For the first control, I generated random nucleotide sequences that are then treated as miRNA sequences. If false positives are unlikely, my predictions should not return probable candidates for these artificial sequences. Finally, I created two sets of real miRNA sequences to repeat this test. The first set comprises miRNAs that are expected to be restricted to the genus *Drosophila*. The second set consists of miRNAs that can only be found in mammals. I expect miRNAs from both of these sets to be absent from the genomes of *X. bocki* and *N. vectensis* and therefore be good candidates to test for false positive predictions of my miRNA prediction pipeline.

5.4.1. MiRNA prediction using simulated data

Simulating miRNA sequences and families

I use simulated data to test the prediction pipeline, gauging the failure rate against arbitrary sequences. Generating and testing random sequences is useful as they are extremely unlikely to exist as real miRNAs within the genomes tested. However, due to

the short length of mature miRNAs (∼20 nucleotides) it is statistically more likely to encounter a similar sequence by chance compared to larger sequences such as protein coding genes. I use simulated data to control for false positives at different conservation thresholds.

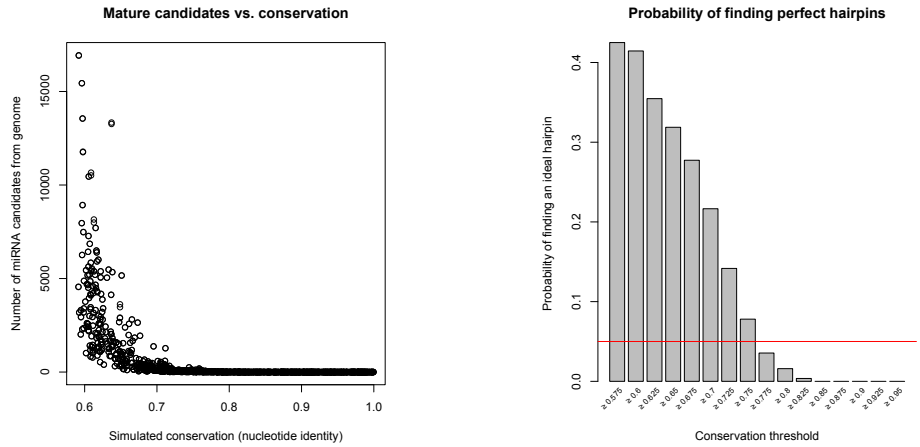
The simplest test is a search for perfectly conserved mature miRNA sequences. It is statistically highly unlikely to find a specific arbitrary sequence within the tested genomes. Assuming a random distribution of nucleotides the chance to generate a mature miRNA perfectly matching a given genome stretch of 22 nucleotides is $(\frac{1}{4})^{22} <$ 1e-13. I tested the false positive rate for perfectly matching miRNAs by creating 10,000 random nucleotide sequences of length 22. I then scanned the whole genome of *Xenoturbella bocki* for instances of these sequences. I also repeated this for another 10,000 samples where each sample's total amount of guanine and cytosine was matched to the GC-content of the *X. bocki* genome (42% G+C ±5%). In both scenarios I did not find a single perfect match.

Mature miRNA sequences in animals are not perfectly conserved. I accounted for this by testing the false positive rate at lower conservation levels. I generated 1,000 random nucleotide sequences (length between 21 and 24, again matched to the GC-content of *X. bocki*). From these sequences I constructed simulated miRNA "pseudo-families". For each pseudo-family I used the initially generated sequence itself as reference sequence and its length as the maximum length to find mature miRNA candidates. For the seed sequence I used nucleotides 2 to 7 from the generated sequence. I then assigned a conservation threshold which was randomly drawn from a uniform distribution between [0.591, 1] (the lower bound is based on minimal conservation found in bilaterian miRNA families). I also gave each pseudo-family a maximum hairpin length (77 to 142 nucleotides, again based on the hairpins observed in real bilaterian miRNA families). I used each of the generated families twice, once for each possible location (5' or 3' strand) of the mature miRNA on the pre-miRNA. This allows me to check the false positive rate for 2000 different combinations of miRNA family characteristics.

Mature miRNA conservation thresholds influence the number of inferred potential miRNA candidates. As expected, I observed many more candidates for the generated sequences when the conservation rate was low (figures 5.6a and 5.7a). In the *X. bocki* genome my generated pseudo-families produced 1,222,876 candidates for mature miRNA sequences. The vast majority of candidates (95%) had a conservation level of less than

69%. The highest conservation for which I was able to find a candidate was 86.86% nucleotide identity. But conservation thresholds as low as 75.03% in some cases resulted in no candidates inferred. I received similar proportions from the genome of *N. vectensis*.

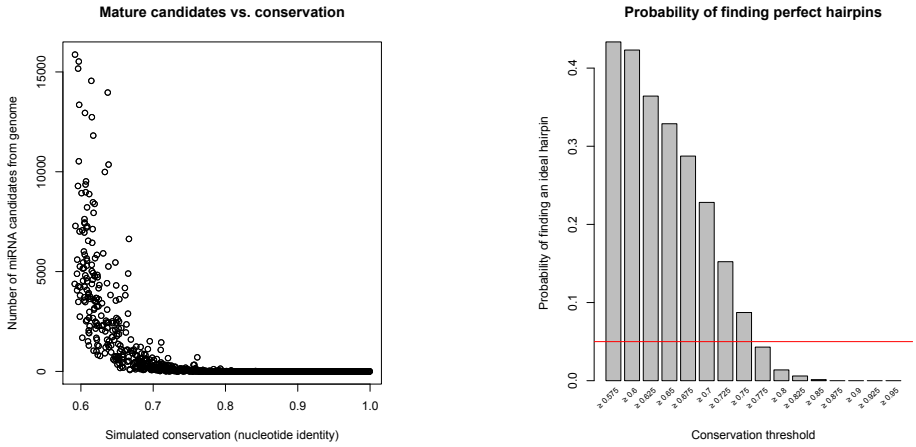
Mature miRNA candidates discovered by sequence similarity alone are not necessarily part of a pre-miRNA that is able to fold into a hairpin structure. Without a viable hairpin my pipeline rejects any falsely identified mature miRNA candidate. The more potential candidates per family, the more likely it is that we find a viable hairpin by chance. It is important to find a conservation threshold at which this possibility is reasonably low. In order to demonstrate this, I created subsets of the pseudo-families according to their assigned conservation. Each subset decreases the conservation by 5%, going from pseudo-families with conservation $\geq$ 95%, $\geq$ 90% and so on ending with a set encompassing all pseudo-families. For each subset I computed the number of pseudo-families for which my prediction pipeline found a perfect hairpin candidate. This proportion represents the amount of false positives at each conservation threshold (figures 5.6b and 5.7b).



(a) Number of candidates at different conservation levels.

(b) Probability of finding ideal hairpins at set conservation thresholds.

Figure 5.6.: Results using simulated miRNA "pseudo-families" in *X. bocki* (genome size: 120Mb) show high correlation between level of conservation (left) and number of mature miRNA candidates. The probability to erroneously identify viable hairpins (right) drops below 5% (red line) at conservation levels of 77.5% and higher.



(a) Number of candidates at different conservation levels.

(b) Probability of finding ideal hairpins at set conservation thresholds.

Figure 5.7.: Results using simulated miRNA "pseudo-families" in *N. vectensis* (genome size: 356Mb) confirm findings in *X. bocki* with slightly increased probability of erroneously identifying viable hairpins (right), likely to be caused by an increased genome size.

Finding a suitable threshold to use is a trade-off between two extremes: setting the threshold too low will cause an increase in the identification of random successes (more false positives) that will put true results into question; setting the threshold too high might increase the validity of the results found, at the cost of potentially excluding true results that fall short of the cutoff (more false negatives). A commonly accepted threshold for scientific methods is a failure rate of 5% (e.g. significance levels expressed by *p*-values). In my simulation I found that for a conservation threshold of at least 77% nucleotide identity, about 3.9% (4.6% in in the about three times larger genome of *N. vectensis*) of the simulated families produce at least one ideal hairpin. This implies that any given predicted candidate that shares 77% or more nucleotide identity with a true miRNA is unlikely to be a false positive finding.

5.4.2. Negative controls using species restricted miRNA data

Additionally to the simulated data we wanted to see if there is a probability of erroneously inferring miRNA candidates using real miRNA sequences. Akin to simulated sequences

these miRNAs should not exist in the genomes of interest. I chose two examples for miRNA families which we deem extremely unlikely to be present in Xenacoelomorpha and Cnidaria. The first set is based on miRNA families which can only be found in the genus *Drosophila* (henceforth referred to as "fly families"). The second set is based on miRNAs that have been found in *H. sapiens*, *M. musculus* and at least 8 more mammalian species, but not outside the Mammalia. The condition for specificity is based on the miRNAs' representation within miRBase.

Negative control using fly families

As an alternative test to randomly generated miRNA sequences I used fly specific families as a negative control. This allows me to search for candidates using real existing animal miRNA families. The number of families for which I predict ideal hairpins is an estimate of the rate of false positives my pipeline generates.

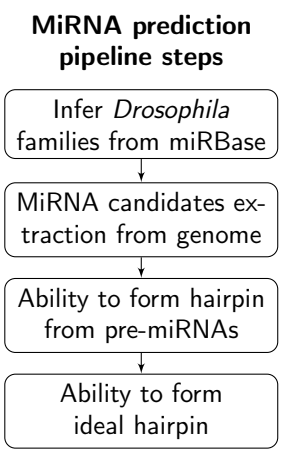
I inferred miRNA families restricted to the genus *Drosophila* from miRBase. MiRBase contains miRNA information from 12 *Drosophila* species. The model organism *Drosophila melanogaster* is the most extensively studied and is represented by 258 precursor miRNA sequences and 469 mature miRNA sequences. Least studied is *Drosophila mojavensis* with 71 sequences for both pre-miRNA and mature miRNA sequences. First, I grouped all mature sequences in miRBase into miRNA families based on their name/number (e.g. *let-7*, *mir-1*). The next step filters for miRNA families specific to *Drosophila*: I removed all families that are also present outside the genus. To increase the validity of a miRNA family, I removed all families that were present in only one *Drosophila* species. This last step allows me to compute the conservation of the miRNAs between individual *Drosophila* species.

As a result I inferred the presence of 49 miRNA families that only exist in the *Drosophila* genus (table 5.2). 10 miRNA families were present in all or all but one species, but most of the miRNAs have only been reported for 2-4 species. For this set I did not require the ancestor to all *Drosophila* species to contain all of these miRNAs, as all of them are expected to not be present in *X. bocki* and *N. vectensis* independent of their origin.

Lower thresholds result in higher number of false positives. As expected, setting the conservation thresholds too low will increase the number of mature miRNA candidates which in turn increases the chance to find a perfect hairpin amongst the corresponding

Name	Strand	Seed	Conservation	$n_{species}$
miR-3	3p	CACUGG	1.000	12
miR-4	3p	UAAAGC	1.000	12
miR-5	5p	AAGGAA	1.000	12
miR-6	3p	AUCACA	1.000	12
miR-280	5p	GUAUUU	1.000	11
miR-284	3p	AAGUCA	0.931	11
miR-287	3p	GUGUUG	0.952	12
miR-288	3p	UUCAUG	0.957	12
miR-289	5p	AAAUAU	1.000	8
miR-311	3p	AUUGCA	0.810	12
miR-313	5p	GCUGCG	0.500	2
miR-314	3p	AUUCGA	1.000	12
miR-955	5p	AUCGUG	1.000	2
miR-956	3p	UUCGAG	1.000	2
miR-958	3p	GAGAUU	1.000	4
miR-959	3p	UGUCAU	0.909	3
miR-960	5p	GAGUAU	0.833	2
miR-961	5p	UUGAUC	0.909	2
miR-961	3p	UCGUUU	0.955	2
miR-962	5p	UAAGGU	0.826	3
miR-963	5p	ACAAGG	0.840	3
miR-964	5p	UAGAAU	0.455	4
miR-967	5p	GAGAUA	0.952	2
miR-968	5p	AAGUAG	0.875	4
miR-969	5p	AGUUCC	0.905	4
miR-973	3p	UCUGUU	0.810	2
miR-974	5p	AGCGAG	0.864	3
miR-975	5p	UAAACA	0.727	3
miR-976	3p	UGGAUU	0.864	3
miR-977	3p	GAGAUA	0.773	3
miR-978	3p	GUCCAG	0.955	2
miR-983	3p	UUAGGU	0.696	2
miR-986	5p	CUCGAA	0.864	4
miR-987	5p	AAAGUA	0.960	4
miR-991	3p	UAAAGU	0.909	2
miR-992	3p	GUACAC	0.591	2
miR-994	5p	UAAGGA	0.955	3
miR-1001	5p	GGGUAA	0.913	2
miR-1002	5p	UAAGUA	0.875	4
miR-1003	3p	CUCACA	0.818	3
miR-1005	3p	CUGGAA	1.000	2
miR-1007	3p	AAGCUC	0.957	2
miR-1010	3p	UUCACC	1.000	4
miR-1011	3p	UAUUGG	1.000	2
miR-1012	5p	UGGGUA	0.955	2
miR-1013	3p	UAAAAG	0.870	2
miR-1017	3p	AAAGCU	0.909	2
miR-2494	3p	UCCCAG	1.000	2
miR-2535	3p	CUCACG	0.864	2

Table 5.2.: Results of *Drosophila* family inference from miRBase: miRNAs had to be present in at least 2 *Drosophila* species and not outside the genus, conservation between species for individual families ranges from 45.5% nucleotide identity to perfect conservation



	Xenoturbella bocki					*Nematostella vectensis*				
	0.65	0.70	0.75	0.80	0.85	0.65	0.70	0.75	0.80	0.85
Infer *Drosophila* families from miRBase	49					49				
MiRNA candidates extraction from genome	49	48		34	14	49			38	22
Ability to form hairpin from pre-miRNAs	49	48	47	26	7	49		48	33	8
Ability to form ideal hairpin	49	47	36	9	4	49		37	17	1

Figure 5.8.: Survival of miRNA families specific to *Drosophila* at each stage of the prediction pipeline (rows) using different thresholds (2nd row, each column shows the results for the given threshold). Thresholds are expressed as nucleotide identity between mature miRNA candidates and reference sequences, numbers represent families that were kept after each step of the pipeline.

pre-miRNAs (fig. 5.8). At 65% conservation threshold, I am able to find viable candidates for all *Drosophila* families in both *X. bocki* and *N. vectensis*. The first noticeable drop in false positives occurs between conservation thresholds of 75% and 80%. This confirms the findings from generated data, where I established a useful cutoff at 77% nucleotide identity. However, unlike with the generated pseudo-families, a large proportion of false positives (18-35%) persists at the higher conservation threshold of 80%. While this number appears to be high, only 2 out of 9 of the best candidates in *X. bocki* and 3 out of 17 in *N. vectensis* have conservation rates above their respective family's minimum conservation threshold.

Negative control using mammalian families

As a second set of miRNA families to test for false negatives I filtered miRBase for miRNAs that occur only in mammals. To decrease the set of families to a manageable size, I filtered for all families that were found in at least 10 mammalian species and contain representative sequences from *Homo sapiens* and *Mus musculus*. As a result of extensive sequencing for these two model organisms, most miRNA families contained sequencing from both strands of the pre-miRNA hairpin structures. For these families I

MiRNA prediction pipeline steps	*Xenoturbella bocki*					*Nematostella vectensis*				
	0.65	0.70	0.75	0.80	0.85	0.65	0.70	0.75	0.80	0.85
Infer mammal families from miRBase	68					68				
MiRNA candidates extraction from genome	67	66	65	42	15	67	65		56	27
Ability to form hairpin from pre-miRNAs	67	65	62	33	12	67	65		51	17
Ability to form ideal hairpin	67	62	50	15	3	66	64	56	24	6

Figure 5.9.: Survival of miRNA families specific to mammals at each stage of the prediction pipeline (rows) using different thresholds (2nd row, each column shows the results for the given threshold). Thresholds are expressed as nucleotide identity between mature miRNA candidates and reference sequences, numbers represent families that were kept after each step of the pipeline.

manually checked the expression profile on the miRBase website to identify the acting strand. I removed families for which the strand differed between the two model organisms or where expression levels were similar for both strands in both species. The resulting set comprises 68 miRNA families (table 5.3).

False positive rate for mammalian specific families is comparable to fly specific families. At low thresholds (less than 75%) I am able to find potential mature miRNA candidates that are able to form viable hairpins (fig. 5.9). Akin to the negative control using fly miRNAs and again confirming the results from simulated data, I find a substantial reduction in predicted candidates between 75% and 80% conservation of the mature miRNA candidate. 22-35% of all mammalian families still yield viable miRNA candidates and ideal hairpins at 80% conservation rate. Only a few of these families (3 in *X. bocki* and 1 in *N. vectensis*) are above their families' respective minimum conservation threshold.

Name	Strand	Seed	Conservation	$n_{species}$
miR-105	5p	UGCUCA	0.783	13
miR-127	3p	UCGGAU	0.917	19
miR-134	5p	GUGACU	0.955	15
miR-136	5p	CUCCAU	0.957	16
miR-149	5p	CUGGCU	1.000	13
miR-181d	5p	ACAUUC	0.958	14
miR-185	5p	GGAGAG	0.957	13
miR-186	5p	AAAGAA	0.957	17
miR-188	5p	AUCCCU	0.913	16
miR-208b	3p	AUAAGA	0.955	12
miR-224	5p	AAGUCA	0.913	14
miR-296	3p	AGGGUU	0.864	13
miR-324	5p	GCAUCC	1.000	13
miR-326	3p	CUCUGG	0.952	13
miR-328	3p	UGGCCC	1.000	12
miR-330	3p	CAAAGC	0.880	11
miR-331	3p	CCCCUG	0.952	15
miR-335	5p	CAAGAG	1.000	14
miR-339	5p	CCCUGU	1.000	8
miR-340	5p	UAUAAA	0.682	13
miR-342	3p	CUCACA	0.880	15
miR-345	5p	GCUGAC	0.773	12
miR-346	5p	GUCUGC	0.864	11
miR-361	5p	UAUCAG	1.000	14
miR-369	3p	AUAAUA	0.952	14
miR-370	3p	CCUGCU	0.917	12
miR-374b	5p	UAUAAU	0.957	12
miR-376a	3p	UAGAGG	0.952	13
miR-376b	3p	UCAUAG	0.818	13
miR-377	3p	AUCACA	0.870	12
miR-379	5p	GGUAGA	0.840	14
miR-380	3p	UGGUCC	0.857	12
miR-381	3p	UAUACA	0.952	15
miR-382	5p	AAGUUG	0.955	12
miR-409	3p	GAAUGU	0.917	12
miR-410	3p	AUAUAA	1.000	13
miR-411	5p	UAGUAG	0.625	11
miR-412	5p	UGGUCG	0.550	8
miR-421	3p	UCAACA	0.870	11
miR-423	5p	GAGGGG	0.958	12
miR-433	3p	UCAUGA	1.000	12
miR-450a	5p	UUUUGC	0.955	13
miR-450b	5p	UUUGCA	0.773	11
miR-452	5p	UGUUUG	0.864	10
miR-484	5p	CAGGCU	1.000	8
miR-485	5p	GAGGCU	0.957	11
miR-487b	3p	AUCGUA	1.000	12
miR-488	3p	UGAAAG	0.952	8
miR-491	3p	UUAUGC	1.000	8
miR-494	3p	GAAACA	0.957	13
miR-495	3p	AACAAA	0.957	13
miR-503	5p	AGCAGC	0.950	11
miR-504	5p	GACCCU	0.880	15
miR-532	5p	AUGCCU	0.955	14
miR-542	3p	GUGACA	0.957	12
miR-543	3p	AACAUU	0.957	12
miR-582	5p	UACAGU	0.955	11
miR-592	5p	UUGUGU	0.957	13
miR-615	3p	CCGAGC	0.955	8
miR-652	3p	AAUGGC	0.875	13
miR-653	5p	UGUUGA	0.810	9
miR-671	5p	GGAAGC	0.958	10
miR-676	3p	AGGUUG	0.818	9
miR-708	5p	AGGAGC	1.000	13
miR-758	3p	UUGUGA	0.955	12
miR-874	3p	UGCCCU	0.917	14
miR-876	5p	GGAUUU	0.905	9
miR-1249	3p	ACGCCC	0.957	13

Table 5.3.: Results of mammalian family inference from miRBase: miRNAs had to be present in at least 10 species (must include *H. sapiens* and *M. musculus*) and not outside the class, conservation between species for individual families ranges from 55% nucleotide identity to perfect conservation

5.4.3. Comparison between negative controls

Negative controls for my miRNA prediction pipeline showed a higher false positive rate when using real miRNA sequences compared to generated pseudo-sequences. It is important to investigate, if this difference is statistical significant and, if so, find the underlying causes. I found that predictions of candidates for real but supposedly absent miRNA families returned more mature miRNA candidates as well as ideal hairpin structures (fig. 5.10). This was consistent across different conservation thresholds.

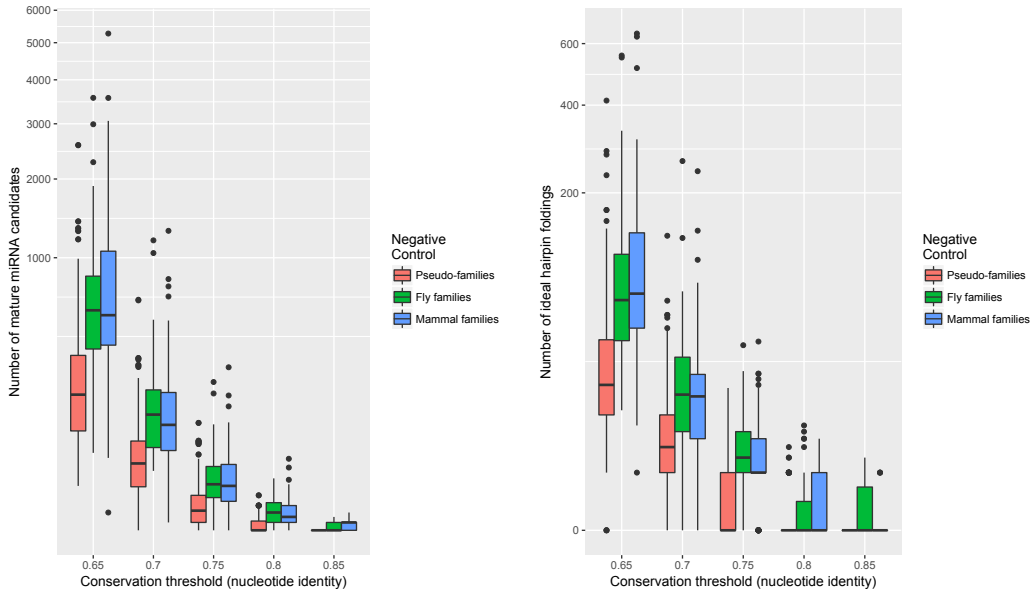


Figure 5.10.: *X. bocki* results of negative controls for miRNA prediction shows that number of predicted candidates is consistently higher for real miRNA sequences, left - number of potential mature miRNA candidates, right - number of predicted ideal hairpins based on mature candidates.

I used a z-test to compare if the increase in predicted candidates is significant. For this test, I will compare the underlying probabilities of predicting miRNA candidates for the given sets of miRNA families. My null hypothesis states that the observed predictions do not differ significantly between tested sets of miRNA families. For this test, I account for all viable ideal hairpins above a conservation threshold of 77% nucleotide identity across all miRNA families tested (pseudo, fly and mammal). For each set I need the number of all miRNA/species combinations (n_i) and the number of observed predictions (m_i) to calculate the observed probability ($\widehat{p_i}$) of predicting a viable miRNA candidate.

For pseudo-families:	
Potential miRNA/species combinations	$n_1 = 2000 * 4 = 8000$
Predicted miRNA candidates	$m_1 = 688$
Observed probability	$\widehat{p_1} = \frac{m_1}{n_1} = \frac{688}{8000} = 0.086$

For fly families:	**For mammalian families:**
$n_2 = 49 * 4 = 196$	$n_3 = 68 * 4 = 272$
$m_2 = 34$	$m_3 = 43$
$\widehat{p_2} = \frac{m_2}{n_2} = \frac{34}{196} \approx 0.173$	$\widehat{p_3} = \frac{m_3}{n_3} = \frac{43}{272} \approx 0.158$
$z_{12} \approx -3.213$	$z_{13} \approx -3.227$
p-value < 0.0023	p-value < 0.0022

Table 5.4.: Results of z-test to compare proportions of predicted miRNA candidates between different sets of miRNA families: finding viable pre-miRNA candidates for real, but supposedly absent miRNA families in *X. bocki* is significantly higher than finding viable candidates for pseudo-families.

The z-test uses the z-statistic (equation 5.1), i.e. the difference between the observed probabilities factoring in the estimated variance ($\widehat{\sigma}_p$) between the tested sets. The z-statistic follows a standard normal distribution, i.e. the significance can be calculated by computing the probability of z_{ij} given my observations (equation 5.2). This test shows that both real miRNA sets yield a significantly higher proportion of predicted miRNAs (table 5.4). I also calculated that the difference between fly and mammalian families is not significant ($p < 0.362$).

$$z_{ij} = \frac{\widehat{p_i} - \widehat{p_j}}{\widehat{\sigma}_p}, \text{ with } \widehat{\sigma}_p = \sqrt{\frac{\widehat{p_i}(1-\widehat{p_i})}{n_i} + \frac{\widehat{p_j}(1-\widehat{p_j})}{n_j}} \tag{5.1}$$

$$p\text{-value} = 2 \cdot Pr(Z > |z_{ij}|), \text{ with } Z \text{ following the standard normal distribution} \tag{5.2}$$

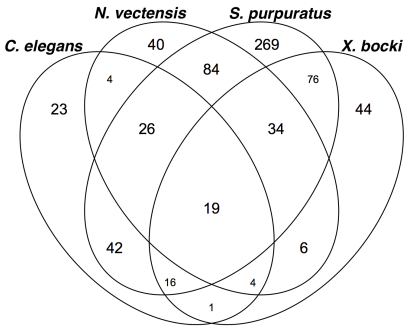
Based on these differences, I looked for common patterns of false positive findings by comparing the proportions of families erroneously identified in more than one species. For all sets I used the 77% conservation threshold, which resulted in less than 5% false positive findings when I used pseudo-families. The proportion of families for which I predicted viable hairpins in several species differed between generated and real miRNA

sequences. For pseudo-families I found that 4.95% (99 families) of all pseudo-families were predicted in 3 or more species (fig. 5.11a). This proportion is higher than what would be expected: given a probability of 0.1 to successfully predict a miRNA candidate regardless of the given family, the estimated number of candidates consistently found in at least 3 species is approximately $8000 \cdot 0.0037 \approx 30$ (equation 5.3). The most likely explanation for this is the fact that genomes are not independent of each other, i.e. predicting a candidate in one species increases the (a-priori) chance of finding a similar sequence in another species due to their shared evolutionary history. However, the proportion of consistently predicted miRNA candidates was much higher in the species restricted families (fig. 5.11) with 18.4% in fly families (9 out of 49 families) and 22.1% in mammalian families (15 out of 68 families). Another z-test shows that this increase is also significant: $p < 0.0328$ for fly families and $p < 0.0035$ for mammalian families.

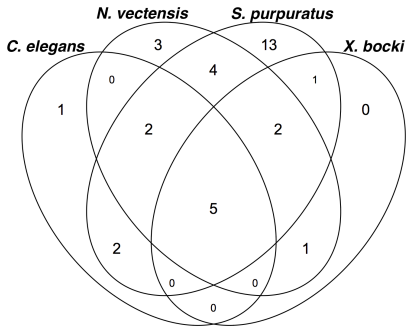
$$\begin{aligned} & p = 0.1, \text{ probability of random prediction} \\ P(n \geq 3) = P(n = 3) + P(n = 4) &= \binom{4}{3} p^3(1-p) + p^4 = 0.0037 \end{aligned} \tag{5.3}$$

5.5. Discussion

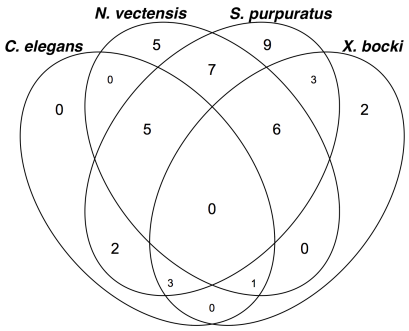
My goal in this chapter was to re-evaluate the apparent lack of certain bilaterian miRNAs in the Xenacoelomorpha. This lack of bilateral characters had been used as support for their phylogenetic position as sister to all remaining bilaterians [Sempere et al., 2006, 2007, Wheeler et al., 2009]. This hypothesis would group protostomes and deuterostomes into a clade called "Nephrozoa" (common occurrence of filtration systems) [Jondelius et al., 2002] which share a common ancestor with the Xenacoelomorpha. The miRNAs conserved amongst nephrozoans but absent from the xenacoelomorphs would have evolved after the proposed divergence from their common ancestor. However, there is already evidence that some of these bilaterian or "nephrozoan" miRNAs initially thought to be absent from Xenacoelomorpha do exist in some xenacoelomorph species. Philippe et al. [2011] sequenced many of these previously missing bilaterian miRNAs in the acoel *H. miamia* and even more in *X. bocki*. The results of my detection pipeline not only confirmed these bilaterian miRNAs to be present within *X. bocki*,



(a) Number of predicted hairpins for pseudo-families in respective species.



(b) Number of predicted hairpins for fly families in respective species.



(c) Number of predicted hairpins for mammalian families in respective species.

Figure 5.11.: Results of negative controls for miRNA prediction, intersections of Venn diagrams show number of families for which a viable hairpin has been predicted in more than one species.

but also found 4 previously undetected bilaterian miRNAs. Unfortunately, I was unable to extend my investigations to other members of Xenacoelomorpha as we do not have recent small RNA sequence data from other xenacoelomorph species. My discoveries nevertheless encouraged me to develop a method that would be able to predict miRNA candidates from genomic data without the need for RNA sequencing.

Predicting mature miRNA candidates from genomic data is feasible. Philippe et al. [2011] found genomic evidence for 4 miRNAs in *X. bocki*. They were unable to confirm these miRNAs from their small RNA sequence dataset. This absence from the RNA transcripts could have been caused by a low level of expression that was excluded from sequencing. The level of expression of miRNAs is affected by developmental and temporal changes. In our newly sequenced small RNA set I detected all 4 of these predicted miRNAs, confirming their presence in *X. bocki*. I predicted mature candidates for an additional 4 miRNA families to be present in the genome of *X. bocki*, which have never been sequenced or predicted so far (table 5.1). All of these putative miRNAs would be able to form viable hairpin structures after expression and have conservation above their respective thresholds. Future studies will show, if these predictions can be confirmed. Furthermore, I predicted viable candidates for an additional 5 of the 36 bilaterian miRNA families tested in this study, but none of these candidates passed the family thresholds and I did not find any of them expressed in the small transcripts.

Stringency of thresholds needs careful consideration. For each miRNA family tested I establish a threshold using member sequences inferred from miRBase. The sequence similarity between the least similar member sequences is used as the conservation threshold for a given family. My detection pipeline uses these thresholds to reject mature miRNA candidates found in the small RNA transcripts. This could be seen as a very conservative approach, especially if we consider the ongoing debate about the relation between xenacoelomorphs and the species that are represented in the bilaterian miRNA families. MiRNA candidates that fall short of the conservation threshold could still be viable and related, but all member sequences reported so far had a higher conservation rate. This is highlighted by one extreme example of my pipeline rejecting a putative miRNA candidate. The found RNA sequence for *mir-153* showed 95.3% similarity with the family's reference sequence, but was excluded due to the 95.5% conservation within the family. For high conservation rates, taking theses inferred thresholds is useful to avoid false positives. The downside is an increased false negative rate, demonstrated by the fact

that I failed to detect miRNAs that had previously been sequenced. I controlled for this on a case by case basis, but future development of the prediction pipeline should include the possibility for automatic cross referencing of prediction results with the RNA data, if available.

MiRNA families present in *X. bocki* also have viable miRNA candidates predicted from acoel genomes. For all families that I detected in *X. bocki* I was able to detect or predict viable candidates in the two acoel species included in this study, *S. roscoffensis* and *P. rubra*. The combined information about presence in RNA and DNA from across the Xenacoelomorpha increases my confidence that these bilaterian miRNAs are present and that their previously reported lack could be an artefact of sparse taxon sampling or insensitive sequencing approaches.

The question about absent bilaterian miRNAs in Xenacoelomorpha remains open. I was able to identify or predict viable hairpins for all of the 36 bilaterian miRNA families tested in *X. bocki*, *S. roscoffensis* and *P. rubra*. I did not identify small transcripts that would verify the existence of 9 families in any of the xenacoel species, but predicted viable sequences from all of their genomes. However, not all putative miRNA sequences pass their families' respective similarity thresholds. This could be a sign, that these predictions are false positives or that these sequences have a higher divergence in Xenacoelomorpha (as shown above with the originally failed identification of a *mir-153* sequence). 4 miRNA families had viable predictions in all xenacoel species. The confirmation of these predictions in several species increases my confidence in the presence of these miRNA families in Xenacoelomorpha.

I encourage the inclusion of more xenacoel species to improve our understanding of miRNA presence and absence in Xenacoelomorpha. As of writing this thesis, Nemertodermatida have not been studied for miRNA presence. If bilaterian miRNAs are truly absent from acoel species, it is not clear if this is true for all of Acoelomorpha. I would also like to emphasise the key position that *X. bocki* holds among the Xenacoelomorpha. As sister to Acoela, a true absence in *X. bocki* and Acoela implies the absence from all of the xenacoelomorph species. This interpretation relies on our data from *X. bocki* to not only be complete, but also representative of all of the Xenoturbellida. I encourage the study of the other recently identified xenoturbellids [Rouse et al., 2016, Nakano et al., 2017] to curb the potential for misinterpretation caused by a lack of sampling in this clade.

MiRNA prediction requires careful testing using negative controls. I have shown that the absence of miRNA candidates from the transcript data is not evidence of absence of the miRNA family from the organism. However, the predictions from genomic data alone could be seen as problematic, due to the probability for predicting false positives, i.e. candidates that are similar to miRNA and pre-miRNA sequences by chance. I devised several negative controls to test for the chance of falsely predicting non-miRNA sequences to be viable candidates.

Generated pseudo-families show low false positive predictions. I generated random nucleotide sequences as representatives of my pseudo miRNA families. These nonsense data should not be part of a real genome and it should be unlikely for me to find similar sequences that also fulfil the needed criteria to work as miRNAs. If I am able to find viable miRNA candidates for these random sequences, it would suggest that a large proportion of my predictions could be false and that my true predictions are a product of chance rather than the results of an informed search strategy. My method was successful in rejecting false positive candidates. At a conservation threshold of 77% sequence similarity between a pseudo miRNA and a found candidate only 3.9% of my pseudo-families yielded a false positive result using our *X. bocki* genome (size: 120Mb) as an example genome. As expected, the larger genome of *N. vectensis* (size: 356Mb) generated more false positives at the same threshold, but this rate remained reasonably low at 4.6% false positive predictions. These results increase my confidence that miRNA candidates predicted from genome data are less likely to be actually absent and only found by random chance.

Negative controls using real miRNA sequences show higher false positive rates. I extracted two sets of miRNA families from miRBase, supposedly specific to *Drosophila* and to mammals respectively, to estimate the false positive rate using real miRNA sequences. I expect these miRNA families to be absent from both of my example genomes due to their apparent taxonomic restriction. However, unlike the results for simulated data, I generated a substantially larger proportion of false positives. As an example, even at a conservation threshold of 80%, predictions of mammal miRNA families in *N. vectensis* returned positive results for $\sim$35% of the tested families. If we assume these miRNAs to be absent, then their behaviour is not comparable with randomly generated data. This would show that the composition of real miRNAs is not random. Alternatively, these miRNAs or sequences that share a high similarity could be present outside the inferred

taxonomic range. Closer investigations are needed to compare these similar sequences between species and if these findings have any bearing on miRNA specificity or sequence variability within members of the same miRNA family. This information could prove useful in adding criteria to more accurately predict the presence of miRNA sequences.

High false positive rate decreases confidence in prediction results. Unfortunately, the results of the negative control using real miRNA data make predictions, especially at lower conservation rates, less reliable. Even if I were to add the restriction of predicting putatively related miRNA candidates in several species, I cannot exclude the possibility that false positives are generated from sequences that are similar between taxa. I experimented with adding additional restrictions to the composition of the miRNA sequence. Bartel [2018] reviewed biochemical experiments about nucleotide motifs that increase the processing efficiency of miRNAs in metazoans. These motifs had already been shown to be much less prevalent in *C. elegans* [Auyeung et al., 2013]. I was also unable to find these motifs in the RNA sequences I identified as miRNA sequences. I therefore decided, it would not be useful to include these criteria as part of my miRNA grading scheme.

Statistical analyses show need for further investigations of real miRNA families. I conducted statistical tests and showed that not only the number of predicted miRNA candidates increased when using real sequences, but also the consistency in predictions across species. This increase was unexpected, as I assumed that sequences supposedly absent from genomes would be predicted akin to generated pseudo-sequences. I was able to show that miRNA like sequences can be found not just sporadically, but consistently in several species. This would hint at features found in miRNA sequences that are common to genomes of species even in the supposed absence of said miRNA sequence. MiRNA sequences or miRNA like structures could exist beyond the current scope captured by miRBase.

Most bilaterian miRNAs conserved in Xenacoelomorpha. Earlier studies depicted acoels as missing a substantial proportion of miRNAs that were shown to be conserved between protostomes and deuterostomes [Sempere et al., 2007, Wheeler et al., 2009]. This fuelled the hypothesis about the xenacoelomorphs as sister to all other bilaterians that diverged before the fixation of other bilaterian miRNAs. Philippe et al. [2011] inferred a position among deuterostomes within the Bilateria. The absence of bilaterian miRNAs would have to be explained by a major loss of the ancestral complement, but their own investigations showed many more miRNAs to be conserved. My own findings

confirm the existence of all reported miRNAs in *X. bocki* and I identified an additional 4 miRNAs that have not been found previously. While individual miRNA prediction results will have to be confirmed, I was able to show that it is highly likely that most of the bilaterian miRNAs in *X. bocki* are also conserved in Acoela.

Bilaterian miRNA complement is not a strong indicator for sister relationship between Xenacoelomorpha and remaining Bilateria. My results continue to refute the notion of xenacoelomorphs having a very small set of bilaterian miRNAs. While we were unable to sequence mature miRNAs for a few bilaterian families, it is not unlikely that these could have been lost in the lineage leading to the xenacoelomorph ancestor. It has been shown that the idea of miRNAs rarely, if ever, being lost is outdated [Fromm et al., 2013]. Closer investigations are necessary to show how the function of miRNAs in other organisms relate to their potential absence in Xenacoelomorpha.

6. General Discussion

The motivation for this PhD project was the investigation of the genomes of the Xenacoelomorpha. This group of marine worms poses questions about phylogenetics, their simple morphology resulting in an ongoing debate about their relation with other members of the animal kingdom and their evolutionary history. The two current hypotheses as to whether they belong outside the Protostomia and Deuterostomia clade as an early off-shoot within Bilateria or as a derived member of Deuterostomia. Both placements have important implications for our understanding of the evolutionary history of the xenacoelomorphs. As sister to all other bilaterians, Xenacoelomorpha would have diverged from the last common ancestor to all Bilateria, informing us about the traits that existed in the common ancestor by evaluating characters shared between xenacoelomorphs and protostomes or deuterostomes. As sister to Ambulacraria (hemichordates and echinoderms), the ancestor of Xenacoelomorpha must have undergone morphological simplification after diverging from a more complex deuterostome ancestor.

In my work I focus on the genetic traits of Xenacoelomopha. Independent of the questions about phylogeny I am interested in the similarities and differences between Xenacoelomorpha and other bilaterians on a genomic level. Here I aimed to establish 2 sources of comparison: groups of orthologous genes (orthogroups) and miRNAs, both specific to Bilateria.

6.1. Orthology inference methods

My first project aimed to infer orthologous genes specific to Bilateria. Currently, there are few publications about bilaterian specific orthologues (e.g. Krämer-Eis et al. [2016]), a paucity which does not seem to reflect the importance of this clade of animals. I noticed several issues in the previous establishment of these orthogroups, which encouraged me to evaluate the factors that influence orthology inference and find methods to scrutinise

the results. My studies have shown that inference methods yield different results and that even congruent results can be scrutinised further.

The publication of new orthology inference methods requires proof of their utility, but ease of use might be more important than applicability. BLAST is probably the most popular program among biologist to search for similar sequences in a wide range of organisms. A big part of its success must be ascribed to its easy to use interface and speed. The BLAST web service allows users to copy and paste a sequence of interest and search the default NCBI database with the click of a button. The results are displayed in order of found similarity in an easily digestible format. I believe that the overwhelming majority of BLAST users are not aware about how the underlying local alignment matching works and how changing parameters could affect the retrieved results beyond the simplified idea of setting a higher *e*-value to allow the capture of "more distantly related" sequences. The treatment of BLAST as a "black box" that takes sequences as input and returns similar sequences as output obscures both advantages and, more importantly, disadvantages of the method. Overall similarity of protein or nucleotide sequences could prove to be superficial when investigating the inferred protein domains and associated functions. Furthermore, using BLAST to infer orthology (e.g. via reciprocal best BLAST hits) is not able to take into account events such as differential gene loss after a gene duplication event, which leads to paralogous genes being inferred as orthologous ("hidden paralogy"). More sophisticated orthology inference methods usually involve a scoring of sequence similarity (often featuring BLAST as part of their pipeline), but add additional steps to ensure the accurate inference of orthologous gene pairs.

The choice of the best available method can be very demanding. Each newly developed method needs show its capabilities compared to established approaches. However, the highlighted improvements in performance might be very specific to the test data set. Even an overall improvement might fail at specific tasks with specific research questions in mind. The published results can only be reliably trusted for the data used in said publication. Applying these new methods to a different set of data, data which may not be as complete or high quality, could result in a worse performance than an approach that had been outperformed on the test data. We are unable to *a priori* determine which method would be best suited for our data and the applied research questions.

Adoption of improved methods is hindered by past use of established methods. Throughout my career and through discussions with peers I have noticed that first choice of methods is usually dictated by existing experience within the research group. Young researchers are faced with the choice of either using a method recommended by group members or being the first to try a new method. The former seems more enticing as questions in applying the method could be answered through discussion with said colleagues. The latter might seem more challenging as it requires the ability and confidence to be able to solve issues on your own and also be willing to justify the use of a new method over established methods. Besides group preferences, method choice can also be informed by approaches being used in the field of study. The implication here would be to generate results that are comparable to previous publications.

Reproducibility and comparability are paramount to make effective choices about method application. To enable reproducibility we need to ensure that all parameters used in our approaches are published among the results. Often I have failed to find out which parameters or which version of a program have been used when reading scientific publications. This prevents readers to recreate the reported outcomes to then compare their own results based on the same framework. It would also be interesting to see how the same analyses would yield different results with newer methods or versions and how this could affect our understanding of the methods and the data investigated. The same is true for updated data such as improved genome quality and annotations. Furthermore, it is difficult to ascertain how different methods perform given a newly produced dataset. The only source of comparison is usually the method's own publication which shows improved performance based on a given dataset. It is not guaranteed that these advantages could be replicated in a different dataset especially if the outcome is yet unknown. Projects such as the Ortholog Benchmarking Webservice (https://orthology.benchmarkservice.org) allow for a more standardised way to compare orthology inference methods, but are only ever able to take into account a very specific dataset with expected outcomes to test against. It would be fascinating to see how applying newer methods to past studies would agree or disagree with the interpreted results, especially if these results caused a change in our biological understanding or stirred controversy.

6.2. Orthologous genes specific to Bilateria

Bilaterian animals comprise most animals living today. Little is known about the origin, morphology and lifestyle of the first bilaterian animal, the "Urbilaterian". Besides the appearance of anterior-posterior and dorso-ventral axes, bilateral symmetry and triploblastic body, it is unclear whether this ancestral organism had larval stages, whether it had a benthic or pelagic life style and what organs evolved after the split from the Cnidaria.

Even less is known about the genetics of the Urbilaterian. We know that Hox genes were involved in its body axis patterning, but it is not clear when and how the bilaterian Hox gene cluster evolved from the Hox genes that already existed in the eumetazoan (Cnidaria + Bilateria) ancestor.

In order to compare Xenacoelomorpha with other bilaterians, I aimed to infer a set of orthologous genes that are shared between the Prostostomia and Deuterostomia. Homologous genes are genes that descended from the same ancestral gene. As a special case, orthologous genes diverged after a speciation event, while paralogous genes diverged after a gene duplication event. Genes orthologous to each other have been observed to have higher sequence similarity compared to paralogous sequences. This retained similarity is interpreted as conservation of the original function. Orthologous genes shared between protostomes and deuterostomes already existed in the common ancestor of these bilaterian clades. Based on this information I could establish key similarities and differences in the genetic makeup of Xenacoelomorpha that may explain their morphology.

I developed several checks to scrutinise orthologous groups in order to achieve a high robustness and reliability even under the use of different methods of inferring orthologous genes. Early on, it became clear that the disagreement between orthology methods cannot be solved by a naïve filtering for groups identically inferred by all methods. I found a way to include groups that were in partial agreement to increase the scope of the first set of orthologous genes that could then be subjected to my reliability tests.

Inference of orthology cannot be reduced to naïvely applying methods and taking the results for granted. I have shown that choosing the parameters for similarity scoring of sequences affects the search of putative homologues. Using the results of a previous publication about deuterostome specific orthologous genes, I showed that increasing the

sensitivity of BLAST (by adjusting the *e*-value) would refute the reported findings. This sensitivity is also the base for the clustering of orthologous genes, i.e. genes that fall below the set similarity threshold are excluded from the clustering. I demonstrated how different orthology inference methods reach different conclusions based on the same input data. Only a small number of orthologous groups were inferred identically by all methods. I amended this by implementing a pairwise checks to allow for groups in partial agreement to be included. It could be argued, that the (partial) agreement of 3 individual inference methods would be enough to secure the validity of these groups of genes. However, I applied 2 more checks to scrutinise these results and to show if these results hold up after extending the search space to include more data. I used the NCBI RefSeq database to fetch additional putative homologous sequences and looked for sequences that might refute the previously inferred clade specificity. Only a fraction of the originally inferred orthologous groups passed my pipeline. This clearly shows that it is imperative to investigate putative groups of orthologous genes before even attempting to interpret their function and importance.

Genes identified as novel, i.e. genes that appear to have evolved *de novo* without relation to other genes, may be particularly unreliable. Using my pipeline, I was able to find several conserved genes for Bilateria, Protostomia, and Deuterostomia, that appeared to have no homologous relationships to genes outside their respective clades. For most of these genes, I was unable to identify protein domain or function, indicating that these genes have not been investigated. This exciting proposition, however, was met with the fact that most of these gene were only shared by 2 or 3 species, which raises questions about the validity of these genes as clade specific and their importance to the clade in general.

Another disappointment was the incongruency with previous findings in Bilateria and Deuterostomia. The differences to previously identified deuterostome specific sets of orthologous gene was expected as I was able to find several issues with the published results. These issues initially led to my implementing steps to verify automatic orthology inference in the first place. Unfortunately, I was also unable to reproduce previously established bilaterian specific clusters of proteins. These results were scrutinised to increase robustness, but the orthology methods I used failed to infer congruent clusters even before I started to integrate and verify the resulting orthologous groups. Krämer-Eis et al. [2016] had used BLAST to establish putative pairwise orthologous sequence

that are specific to Bilateria before they clustered these sequences with OrthoMCL and InParanoid. In my analysis I did filter after computing similarity scores (via DIAMOND) and clustering with different orthology inference methods including OrthoMCL. However, even when focusing exclusively on OrthoMCL's clustering, I was not able to find the same groups of genes identified previously. More research is necessary in how parametrisation of BLAST/DIAMOND and OrthoMCL as well as the inclusion of an increased number of genomes affect the clustering congruency.

It remains challenging to investigate characters over long evolutionary timespans. The focus on individual genes as "keys" to the emergence of a clade might be missing the bigger picture with the currently available methodology and data. The more genes have diverged, due to evolutionary pressure or time, the harder it becomes to accurately infer their relations. Higher divergence requires lower similarity thresholds to capture ancient relations. However, lower thresholds increase the probability of unrelated genes to be inferred as homologous. The sequence similarity of these unrelated genes is a result of chance rather than common evolutionary history and too lenient thresholds would hinder inference methods in discerning these two scenarios.

The potential for errors in orthology inference could lead to arbitrary interpretations of evolutionary changes. I highlighted the issues using different orthology inference approaches and the disagreement with findings of other studies. There is no gold standard to validate the results of inference methods, only reference sets of orthologous genes that may or may not be comparable to specific cases in question. In most cases, I found that the results of orthology inference are accepted as valid, often without additional checks to increase scrutiny. Afterwards these genes are used to find associated functions and pathways in order to interpret these results. I propose investigations in the opposite direction: starting with pathways (or significant changes thereof) and traits absent in outgroups, we need to identify the genes and gene interactions that are involved in the emergence of these characters. Current studies into clade specific genes often rely on significant changes in individual genes or gene sets and rarely take into account what effects these changes had on other non-specific genes and pathways.

6.3. MicroRNA detection and prediction

Little is known about what distinguishes miRNAs from other sequences that form similar structures. MiRNAs are mainly described through their biogenesis pathway and the expression of mature sequences which is higher compared to other sequences of similar length. Perturbation experiments of specific miRNAs such as *let-7* and *lin-4* have shown their involvement in gene regulation [Lee et al., 1993, Wightman et al., 1993, Berezikov et al., 2005, Zhang et al., 2006], but it is unclear what specific characters allow or prevent a short nucleotide sequences to act in this way. MiRNA identification usually involves the reconstruction of the pre-miRNA sequence that is required to fold into a hairpin structure in order to comply with the miRNA pathway. The ability to form a hairpin structure is a necessary qualifier, but Bentwich et al. [2005] have shown that $\sim$11 million potential hairpins exist in the human genome while fewer than 700 human miRNAs have been confirmed so far [Fromm et al., 2015]. Several miRNA detection approaches, including my own, have tried to identify characteristics that distinguish a pre-miRNA from other hairpin structures.

The infancy of miRNA research is reflected by the low identification rates of currently used miRNA detection and prediction methods. With more than 800 citations each, miRDeep2 [Friedländer et al., 2012] and miRseeker [Lai et al., 2003] are the most popular computational miRNA detection methods. Depending on the species, sensitivity of these methods can be as low as 71-75% which shows us that we have yet to identify markers that unite all miRNAs. Alternatively, these results could indicate different "classes" of miRNAs that are defined by separate sets of characteristics. True positive findings for newly detected miRNAs are even more worrying. In *Drosophila* only 20 out of 27 highest scoring miRNAs predicted by miRseeker could be experimentally verified. MiRDeep2's own estimation reports true positive rates as low as 44$\pm$5% for species with more than 10 novel miRNA predictions. These rates might improve once we are able to afford more validation experiments to include lower scoring predictions. For the time being, we are tasked with improving the outcome of the best predictions. The application of more advanced machine learning algorithms might help us to identify which characters or sets of characters are important to predict miRNAs more accurately. Our investigations in the biogenesis, function and location of miRNAs might help us to identify additional restrictions outside the miRNA sequence itself.

6.4. New approaches to find conserved microRNA families

My second project focused on miRNAs conserved across Bilateria. Previous publications have already used a set of miRNAs to investigate conservation in different bilaterians. I used data from miRBase to extend this set for miRNAs shared between protostomes and deuterostomes, but I encountered several issues. I implemented a miRNA identification pipeline to find and evaluate candidates from these miRNA families using genomic and transcriptomic data. For *Xenoturbella bocki* I was able to show a presence of many of the bilaterian miRNAs. I lacked transcriptomic data for acoels and, instead, developed a method to predict miRNA candidates from genomic data alone. These miRNA candidates show that many of the bilaterian miRNAs also exist in the genomes of Acoela.

In my second project I looked at miRNAs that can be found in Xenacoelomorpha. As I am interested in the similarities and differences of xenacoelomorphs and other bilaterians, I specifically studied miRNAs that have been found to be conserved among protostomes and deuterostomes These miRNAs have already existed in the bilaterian ancestor, but previous studies found many to be absent in acoels. MiRNAs are regarded as important genetic markers that have been associated with an increase in structural complexity. Evolution and fixation of new miRNAs have been correlated with the emergence of clades. Due to their high conservation and interpreted importance, loss of miRNAs is seen as rare or unlikely. The absence of many of the bilaterian miRNAs in Acoela was therefore used to support the exclusion of acoelomorphs from the other bilaterian clades.

In a more recent publication some of those missing miRNAs were found in an acoel and *Xenoturbella bocki*. These findings were our motivation to reinvestigate this issue using current sequencing and identification methods.

Unlike other miRNA methods, that try to identify miRNAs using no or little knowledge about related sequences, my goal was to assess the miRNA complement of Xenacoelomorpha specifically compared to miRNAs conserved among other bilaterians. This lead to the development of two methods: a) identification of miRNAs conserved between species based on publicly available data, b) identification of conserved miRNAs in Xenacoelmorpha based on newly sequenced genomic and transcriptomic data. Unfortunately, due to technical difficulties, we were only able to acquire new transcriptomic data for

X. bocki, but not our acoel species. This lead to my developing a third method, which predicts miRNAs based on genomic data alone.

Identification of conserved miRNAs based on published data is cumbersome and requires manual correction. While curated miRNA databases exist, these focus mainly on model organisms or a very restricted taxonomic range. MiRBase is currently the widest ranging and most comprehensive database for miRNA sequences, but during the initial miRNA data acquisition I became aware that miRBase is not very well curated. I have noticed several issues with the description of sequences in miRBase. These range from an inconsistency in following naming conventions (only some of which are historically justified) to misattribution of sequences to miRNA families. While I was able to fetch and identify miRNAs conserved between species of interest automatically, I was forced to re-evaluate the inferred families after finding several details that were obviously misattributed in miRBase. Some families included sequences that did not share the perfectly conserved seed region which is key to be identified as part of a miRNA family. In most cases, I was able to ignore the aberrant sequence without having to exclude the inferred family. However, I showed 3 specific examples, where the excluded sequence was the only representative for a clade in question. This led to the exclusion of the whole family from my analysis, as I was not able to infer the family's conservation based solely on the miRBase data. I can only assume, that these sequences were misannotated or misattributed at a time when there was less knowledge about miRNA characteristics. Rather than presenting a fully automatic pipeline to infer this information, my work provides a semi-automatic way and guidelines on how to retrieve and scrutinise miRNA families for future clades of interest.

I devised a method to find and test the viability of mature miRNA candidates. Based on the information gathered from miRBase I developed a pipeline that identifies mature miRNA candidates, matches them to the genome and extracts a precursor miRNA sequence candidate. I present a novel method to evaluate the folding of this precursor sequence to an idealised hairpin structure to grade potential candidates and to reject those not conforming to certain criteria. This evaluation currently hinges on the usefulness of the characteristics of a hairpin folding structure to determine viable miRNA candidates. Experiments have shown that the processing efficiency of dicer depends on the number of base pairings within the pre-miRNA folding structure. Longer stretches of unpaired regions (loop region and bulges along the stem) decrease processing efficiency

reducing viability of the putative candidate. In an attempt to be conservative, I chose a nearly perfect template to base my grading on. Future research will determine, if this template is too rigid and if more characters can be included, which would improve the grading scheme to more accurately reflect the biology of miRNA processing.

Lastly, I present a method to predict miRNA candidates in the absence of transcriptomic data. Based on my detection pipeline, I developed a method that could predict potential candidates directly from the genome. This pipeline uses the same information about miRNA families of interest and hairpin folding characteristics to extract and evaluate the viability of precursor miRNA candidate sequences. In the absence of transcriptomic data, I needed to find a way to reduce the number of false positive candidates. I successfully used miRNA pseudo-families to demonstrate the specificity of my prediction method. More negative controls using real miRNA sequences, however, revealed a much higher rate of falsely predicted candidate sequences. A closer investigation of the false predictions revealed that real sequences are either more widely distributed across the taxonomy than stated in miRBase or that these sequence feature characteristics that are more widespread than the miRNAs themselves. More research into the structural features of miRNAs is necessary to improve the credibility of currently available and predicted miRNA sequences. Disruption experiments targeting the candidates found will also help to elucidate function and impact of conserved miRNAs across several species.

6.5. MicroRNAs conserved between Xenacoelomorpha and other bilaterians

The existence of miRNAs conserved among protostomes and deuterostomes, but absent from Xenacoelomorpha, was seen as key evidence to place xenacoelomorphs as sister to a clade comprising protostomes and deuterostomes [Sempere et al., 2006, 2007, Wheeler et al., 2009].

New data supports the conservation of many bilaterian miRNAs in *Xenoturbella bocki*. Philippe et al. [2011] sequenced and identified many miRNAs that were thought to be missing from Xenacoelomorpha. With our newly sequenced set of small RNAs from *X. bocki* and my detection pipeline, I was not only able to confirm these miRNAs' existence, but add even more to the list of sequences shared with other bilaterians.

Prediction of miRNA candidates in Acoelomorpha hints at widespread conservation of bilaterian miRNAs among xenacoelomorphs. My predictions of miRNA precursor candidates suggest viable candidates for all of the miRNA families tested, but some fall below either miRNA family thresholds or have a low sequence similarity to the investigated miRNA family. These findings could present false positives in my miRNA prediction which are only similar by chance and do not represent a true presence. Excluding the possibility of incompleteness of investigated genomes, if families will be shown to be absent in acoels, than those miRNA families are likely to have been lost in the branch leading to their ancestor.

A significant absence of bilaterian miRNAs in Xenacoelomorpha cannot be supported and does not inform phylogeny. My studies have shown that past assumptions about the miRNA complement in xenacoelomorphs may have been caused by data incompleteness and lack of species sampling. I support the notion that Xenacoelomorpha share most of the bilaterian miRNAs and may lack only a small subset. Together with recent findings about the frequency of miRNA loss, I do not believe that the presence and absence of bilaterian miRNAs are sufficient to provide support for the phylogenetic relation of xenacoelomorphs as sister to the remaining bilaterians.

6.6. Xenacoelomorpha - current status and outlook

Xenacoemolorpha are a phylogenetic enigma due to their arguably simple morphology. The lack of morphological characters found in other bilaterians such as through-gut and inner organs made a compelling argument about their position as sister to all other bilaterians. This position was supported by the idea that "more complex" characters evolved in the protostomes and deuterostomes after the split from a common ancestor [Westblad, 1949]. Despite that, several morphological characters especially in *X. bocki* were described as similar to more derived bilaterians: cilia structure and the statocyst were likened to those of hemichordates, but the phylogenetic relevance of these characters was disputed. The lack of morphological complexity may not be informative as a distinguishing character. Platyhelminthes, originally placed as sister to protostomes and deuterostomes alongside Xenacoelomorpha, were shown to be part of

the Lophotrochozoa within the Protostomia [Balavoine, 1997, Carranza et al., 1997],. Their morphology had to be reinterpreted as many absent characters had to be a result of secondary loss rather than representing an ancestral state. The lack of observed characters may also be an artefact caused by the lack of histological investigations. Thanks to the investigations of the newly discovered *Xenoturbella japonica* sp. nov. a glandular network has been identified which also exists in acoels and nemertodermatids. Continued research might reveal more features that have not yet been described.

The discussion about the phylogenetic position of Xenacoelomorpha is also a discussion about phylogenetic models and systematic errors. Throughout my literature review I have noticed a peculiarity about phylogenetic analyses involving xenacoelomorphs: studies featuring only acoelomorph species consistently place Acoelomorpha as sister to protostomes and deuterostomes, while studies that only included *X. bocki* (usually excluding acoelomorphs for their incomplete genomes or high evolutionary rate) place it among deuterostomes. Broader phylogenetic analyses that include both taxa have resulted in the currently ongoing dispute about the their true position. The discussion about the correct position also revolves around the proper choice of gene selection and gene models. The Xenambulacraria hypothesis is usually supported by the argument for arguably more sophisticated site-heterogenous models which are said to fit the data better while reducing the impact of systematic errors. As part of one of the groups supporting this hypothesis I have seen the careful examination about which genes are better at reproducing known phylogenies and would therefore be better suited to solve this issue (manuscript currently under review). That being said, I am well aware of the possibility of bias that stems from being part of the debate. Our approach including the chosen methods and models will no doubt have to be re-examined for its robustness in the future.

The recent identification of 5 more species of *Xenoturbella* promises a more in-depth understanding of this enigmatic clade of marine worms. Until now, *Xenoturbella bocki* had a pivotal position among xenacoelomorphs, being the slowest evolving member and sister to all other xenacoelomorphs. These new species will allow us to close gaps caused by the undersampling of the genus *Xenoturbella*. The aforementioned identification of the glandular system could be the first of many more discoveries as a result of comparing individual xenoturbellid species or comparisons of xenoturbellids, besides *X. bocki*, with Acoelomorpha and other Bilateria.

Inclusion of more acoelomorph, especially nemertodermatid, species is also necessary. *X. bocki* is by far the most well studied of the xenacoelomorphs. The increasing efforts throughout history have shown to reveal characters previously thought to be absent in this clade. Furthermore, many findings attributed to Acoelomorpha are based on findings in acoels, as there is only very little data on nemertodermatids. Future studies should aim to increase sampling for this clade, as well.

More investigations of "absent" characters is needed. Xenacoelomorphs seem to lack specific organs for ultrafiltration, which was also used to support a position as sister to bilaterians with ultrafiltration systems ("Nephrozoa", i.e. protostomes and deuterostomes). The inferred position of Xenacoelomorpha as sister to the Ambulacraria within Deuterostomia implies the simplification from a more complex ancestor. In her PhD project, Helen Robertson detected the conservation and co-expression of genes associated with ultrafiltration in *X. bocki*. While these results cannot be used as definitive evidence to support either of the phylogenetic hypotheses, it changes our perception of character absence within our clade of interest. Future studies need to investigate other supposedly absent genes and characters to accurately describe the changes that lead to the Xenacoelomorpha ancestor.

A. Appendix - Genome sources for orthology inference

Adineta vaga
http://www.genoscope.cns.fr/adineta/data/Adineta_vaga.v2.pep.fa.gz

Amphimedon queenslandica
ftp://ftp.ncbi.nlm.nih.gov/genomes/Amphimedon_queenslandica/protein/protein.fa.gz

Branchiostoma floridae
http://www.uniprot.org/proteomes/UP000001554

Caenorhabditis elegans
ftp://ftp.ensembl.org/pub/current_fasta/caenorhabditis_elegans/pep/Caenorhabditis_elegans.pep.all.fa.gz

Callorhinchus milii
ftp://ftp.ncbi.nlm.nih.gov/genomes/Callorhinchus_milii/protein/protein.fa.gz

Capitella teleta
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Capitella_teleta/pep/Capitella_teleta.pep.all.fa.gz

Capsaspora owczarzaki
ftp://ftp.ensemblgenomes.org/pub/protists/current/fasta/protists_ichthyosporea1_collection/capsaspora_owczarzaki_atcc_30864/pep/Capsaspora_owczarzaki_atcc_308

64.C_owczarzaki_V2.31.pep.all.fa.gz
Ciona intestinalis
ftp://ftp.ncbi.nlm.nih.gov/genomes/Ciona_intestinalis/protein/protein.fa.gz

Danio rerio
ftp://ftp.ncbi.nlm.nih.gov/genomes/Danio_rerio/protein/protein.fa.gz

Daphnia pulex
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Daphnia_pulex/pep/Daphnia_pulex.pep.all.fa.gz

Drosophila melanogaster
ftp://ftp.ensembl.org/pub/current_fasta/Drosophila_melanogaster/pep/Drosophila_melanogaster.pep.all.fa.gz

Gallus gallus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Gallus_gallus/protein/protein.fa.gz

Helobdella robusta
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Helobdella_robusta/pep/Helobdella_robusta.pep.all.fa.gz

Homo sapiens
ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/protein/protein.fa.gz

Hydra vulgaris
ftp://ftp.ncbi.nlm.nih.gov/genomes/Hydra_vulgaris/protein/protein.fa.gz

Hypsibius dujardini dujardini
http://badger.bio.ed.ac.uk/H_dujardini/home/download(Makerpredictions)

Ixodes scapularis
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Ixodes_scapularis/pep/Ixo

des_scapularis.pep.all.fa.gz

Latimeria chalumnae
ftp://ftp.ncbi.nlm.nih.gov/genomes/Latimeria_chalumnae/protein/protein.fa.gz

Lepisosteus oculatus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Lepisosteus_oculatus/protein/protein.fa.gz

Lottia gigantea
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Lottia_gigantea/pep/Lottia_gigantea.pep.all.fa.gz

Mnemiopsis leidyi
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Mnemiopsis_leidyi/pep/Mnemiopsis_leidyi.pep.all.fa.gz

Monosiga brevicollis
ftp://ftp.ensemblgenomes.org/pub/protists/current/fasta/protists_choanoflagellida1_collection/Monosiga_brevicollis/pep/Monosiga_brevicollis.pep.all.fa.gz

Mus musculus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Mus_musculus/protein/protein.fa.gz

Nematostella vectensis
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Nematostella_vectensis/pep/Nematostella_vectensis.pep.all.fa.gz

Oncopeltus fasciatus fasciatus
https://www.hgsc.bcm.edu/milkweed-bug-genome-project

Patiria miniata
http://www.echinobase.org/Echinobase/PmDownload

Petromyzon marinus
ftp://ftp.ensembl.org/pub/current_fasta/Petromyzon_marinus/pep/Petromyzon_marinus.pep.all.fa.gz

Priapulus caudatus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Priapulus_caudatus/protein/protein.fa.gz

Ptychodera flava
http://octopus.unit.oist.jp/HEMIDATA/pfl.prot

Rattus norvegicus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Rattus_norvegicus/protein/protein.fa.gz

Romanomermis culicivorax
http://www.nematodes.org/genomes/romanomermis_culicivorax/

Saccoglossus kowalevskii
ftp://ftp.ncbi.nlm.nih.gov/genomes/Saccoglossus_kowalevskii/protein/protein.fa.gz

Salpingoeca rosetta
ftp://ftp.ensemblgenomes.org/pub/protists/current/fasta/protists_choanoflagellida1_collection/Salpingoeca_rosetta/pep/Salpingoeca_rosetta.pep.all.fa.gz

Schmidtea mediterranea
http://smedgd.stowers.org/downloads/#MAKER_annotations_8211_Protein_FASTA_files(SmedSxl4.0prediction)
Strigamia maritima
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Strigamia_maritima/pep/Strigamia_maritima.pep.all.fa.gz

Strongylocentrotus purpuratus
ftp://ftp.ncbi.nlm.nih.gov/genomes/Strongylocentrotus_purpuratus/protein/protein.fa.gz

Tribolium castaneum
ftp://ftp.ncbi.nlm.nih.gov/genomes/Tribolium_castaneum/protein/protein.fa.gz

Trichoplax adhaerens
ftp://ftp.ensemblgenomes.org/pub/metazoa/current/fasta/Trichoplax_adhaerens/pep/Trichoplax_adhaerens.pep.all.fa.gz

Xenopus tropicalis
ftp://ftp.ncbi.nlm.nih.gov/genomes/Xenopus_Silurana_tropicalis/protein/protein.fa.gz

B. Appendix - Scripts published on GitHub

B.1. OrthoMerge pipeline scripts

GitHub repository: `https://github.com/TelfordLab/OrthoMerge`

Python script to merge results from OrthoMCL, OrthoFinder and OrthoInspector:
OrthoMerge.py

Python scripts to validate merged results using monophyly and reciprocal best bidirectional hit tests:
OG_monophyly_test.py
OG_rBBH_test.py

B.2. microRNA detection and prediction scripts

GitHub repository: `https://github.com/TelfordLab/microRNAs`

Python scripts for filtering miRNA families conserved across phyla based on miRBase (`http://mirbase.org/`) sequence information:
filter_miRBase_for_bilaterian_families.py
filter_miRBase_for_drosophila_families.py
filter_miRBase_for_mammalian_families.py

Python script for detecting miRNA candidates using small RNA and genome data based on conserved miRNA family information:

microRNA_detection.py

Python scripts for predicting miRNA candidates using genome data based on conserved miRNA family information after splitting genome sequences into kmers:

split_genome_into_miRNA_kmers.py

microRNA_prediction_from_DNA_kmers.py

D240–D249 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 15 November 2014*
doi: 10.1093/nar/gku1158

The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements

Adrian M. Altenhoff[1,2,3], Nives Škunca[1,2,3], Natasha Glover[1,4,5], Clément-Marie Train[3], Anna Sueki[1], Ivana Piližota[1], Kevin Gori[6], Bartlomiej Tomiczek[1], Steven Müller[1], Henning Redestig[5], Gaston H. Gonnet[2,3] and Christophe Dessimoz[1,6,*]

[1]University College London, Gower Street, London WC1E 6BT, UK, [2]Swiss Institute of Bioinformatics, Universitätstr. 6, 8092 Zurich, Switzerland, [3]ETH Zurich, Computer Science, Universitätstr. 6, 8092 Zurich, Switzerland, [4]Institut National de la Recherche Agronomique (INRA) UMR1095, Genetics, Diversity and Ecophysiology of Cereals, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand, France, [5]Bayer CropScience NV, Technologiepark 38, 9052 Gent, Belgium and [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 22, 2014; Revised October 24, 2014; Accepted October 29, 2014

ABSTRACT

The Orthologous Matrix (OMA) project is a method and associated database inferring evolutionary relationships amongst currently 1706 complete proteomes (i.e. the protein sequence associated for every protein-coding gene in all genomes). In this update article, we present six major new developments in OMA: (i) a new web interface; (ii) Gene Ontology function predictions as part of the OMA pipeline; (iii) better support for plant genomes and in particular homeologs in the wheat genome; (iv) a new synteny viewer providing the genomic context of orthologs; (v) statically computed hierarchical orthologous groups subsets downloadable in OrthoXML format; and (vi) possibility to export parts of the all-against-all computations and to combine them with custom data for 'client-side' orthology prediction. OMA can be accessed through the OMA Browser and various programmatic interfaces at http://omabrowser.org.

INTRODUCTION

The flood of newly sequenced genomes presents a daunting interpretation challenge. Fortunately, the common origin of all living beings implies that many genes are conserved across species—in some cases despite billions of years of intervening evolution. Elucidating evolutionary relationships amongst genes and genomes is thus a key step in the analysis of new data. Sequences that have a common ancestry—homologs—are typically refined into orthologs, which are pairs of genes that started diverging via speciation, and paralogues, which are pairs of genes that started diverging via gene duplication (1,2). This distinction is useful in a broad range of contexts, including multigene phylogenetic inference, propagation of experimental knowledge from model organisms to non-model organisms and the study of gene evolution and adaptation (reviewed in 3,4). The need for orthology inference has led to the development of numerous methods (reviewed in 5) and databases, notably including EggNOG (6), Ensembl Compara (7), Inparanoid (8), MBGD (9), OrthoDB (10), OrthoMCL (11), Panther (12), PhylomeDB (13), Plaza (14) and OMA (15).

The OMA (Orthologous MAtrix) project is a method and database for the inference of orthologs amongst complete proteomes (i.e. the protein sequences associated for every protein-coding gene in all genomes). Initiated in 2004, OMA has undergone 17 major releases, steadily increasing the number of proteomes under consideration from 150 to 1706 across all domains of life. Besides its large scope, the distinctive features of OMA are the high specificity of the inferred orthologs (e.g. 16–19), feature-rich web interface, availability of data in a wide range of formats and interfaces and frequent update schedule of two releases per year.

In this update paper, after providing a brief review of the OMA pipeline, we present major new features recently added to OMA: a new web interface and reorganization, integrated gene ontology function prediction, better support of plant genomes, a synteny viewer depicting orthology relationships in their genomic context, statically computed hierarchical orthologous groups (HOGs) and the possibility to export genomes including all-against-all computations and to combine them with custom genome/transcriptome data.

*To whom correspondence should be addressed. Tel: +44 20 7679 0079; Fax: +44 20 7679 7193; Email: c.dessimoz@ucl.ac.uk

OVERVIEW OF THE OMA INFERENCE PIPELINE

OMA's inference algorithm consists of three main phases:

(i) First, to infer homologous sequences (sequences of common ancestry), we compute all-against-all Smith–Waterman alignments between every sequence and retain significant matches.
(ii) Second, to infer orthologous pairs (the subset of homologs related by speciation events), mutually closest homologs are identified based on evolutionary distances, taking into account distance inference uncertainty and the possibility of hidden paralogy due to differential gene losses (20,21).
(iii) Third, these orthologs are clustered in two different ways, which are useful for different purposes: (a) we identify cliques of orthologous pairs (OMA groups). Because all relations in one OMA group are orthologous, these are useful as marker genes for phylogenetic reconstruction and tend to be highly specific (18); (b) we identify HOGs, groups of genes defined for particular taxonomic ranges and identify all genes that have descended from a common ancestral gene in that taxonomic range (22).

OMA infers evolutionary relationships between genes from protein sequences, using one protein sequence per gene. If multiple splicing variants are possible, the best one in terms of matches with other genomes is selected, which is not necessarily the longest one (15).

NEW WEB INTERFACE WITH BETTER ORGANIZATION

The OMA browser has been reorganised and redesigned to make it user-friendlier. The menu bar provides a consistent and persistent overview of all main functionalities. The documentation and help pages have been restructured and extended. The new 'responsive' layout takes advantage of large contemporary screens whilst also accommodating small screens such as smartphones and tablets. The landing page now provides pointers to introductory explanations for new users and recent announcements for returning users (Figure 1).

GENE ONTOLOGY FUNCTION INFERENCE AS PART OF THE OMA PIPELINE

One key motivation for orthology inference is to computationally predict the roles that genes play in living organisms—e.g. Cellular Component, Molecular Function and Biological Process of the Gene Ontology (23). For many years, Gene Ontology (GO) annotations from the UniProt-GOA database (24) have been linked to all sequences in OMA. Additionally, we now provide inferred annotations based on orthology relationships: within the orthologous groups, we propagate GO annotations across different species.

To infer GO annotations, we start with curated annotations that are based on direct evidence from the literature: GO evidence codes EXP, IDA, IPI, IMP, IGI and IEP (http://geneontology.org/page/guide-go-evidence-codes). We then propagate them across OMA groups—sets of genes for which all members are inferred to be mutually orthologous—as these have been previously shown to be highly coherent in terms of functional annotations (25). Additionally, to avoid over-propagating clade-specific terms (e.g. 'nematode larval development' outside the nematodes), we require that propagated terms be used in at least one literature-based annotation in the clade in question. For example, the OMA group with fingerprint 'VWQCDTP' contains a *Caenorhabditis elegans* gene annotated with the GO term 'nematode larval development' (Figure 2); this term is not appropriate for genes outside of the Nematoda phylum. Therefore, when propagating this GO term to, for example, the poorly annotated *Arabidopsis thaliana* protein within the same OMA group, we only propagate those parent terms of 'nematode larval development' that are known to be associated with plant proteins; in this case, the most specific amongst those is 'post-embryonic development' (Figure 2). Indeed, the propagated annotation complements one of the known annotations for the *A. thaliana* protein, 'embryo sac development'.

Overall, the OMA database now provides 442 376 477 function annotations for 7 947 728 proteins (Figure 3). Amongst the available annotations, most are computationally inferred; our own predictions constitute about 20% of the available annotations.

Function annotations based on OMA orthologs are particularly valuable for proteins for which other computational annotation methods provide no annotations and the available annotations assigned by curators are relatively general and/or sparse. In the most recent OMA release, we provide annotations for 423 983 proteins for which there are no other electronic annotations. For example, at the time of writing the *A. thaliana* protein with UniProt identifier Q8VYZ5 had no electronically inferred GO annotations (evidence code IEA); it had five annotations based on evidence codes ISS or RCA, which are not used in our propagation pipeline; and the annotations from literature-based evidence were 'nucleolus' (IDA), 'rRNA processing' (IMP) and 'embryo sac development' (IMP). Using our OMA annotation pipeline, we assigned new annotations that complement these: for example, we inferred GO terms 'RNA 5'-end processing' and 'endonucleolytic cleavage involved in rRNA processing' that complement the known experimental annotation 'rRNA processing'; we inferred the GO term 'post-embryonic development' that complements the known experimental annotation 'embryo sac development' (Figure 3).

BETTER SUPPORT FOR PLANT GENOMES, INCLUDING HOMEOLOGY IN WHEAT

One research area where comparative genomics can make an important difference is modern crop science. Indeed, plant genomes tend to have highly redundant genomes as a result of their complex history of duplication and hybridisation events. With almost all genes being available in several copies on multiple sub-genomes, the use of comparative genomics is essential in order to map knowledge across different species. Several specialised plant resources already exist—such as Ensembl Plants (26), Gramene (27),

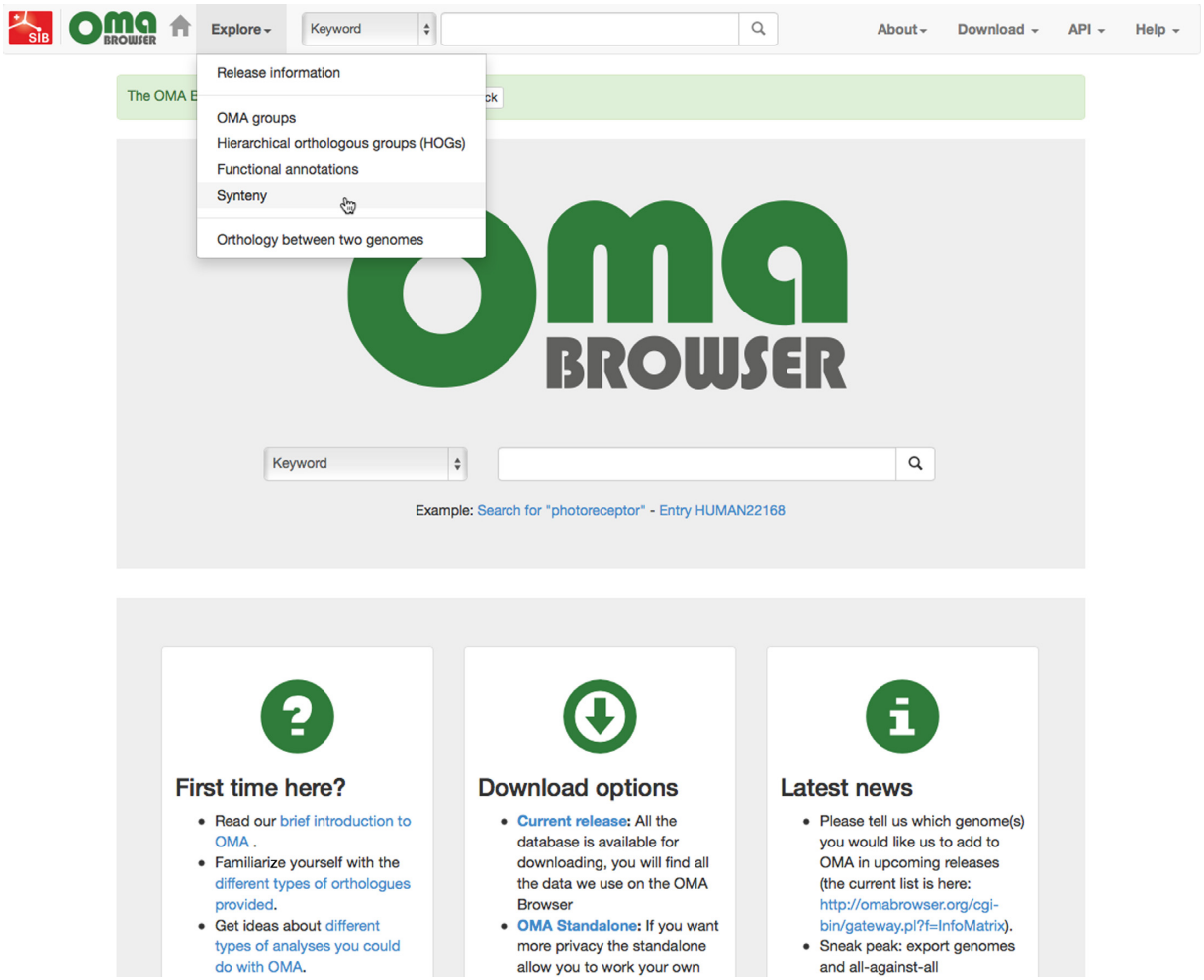


Figure 1. User-centric new design. The website has been redesigned with an emphasis on usability.

Greenphyl (28) and Plaza (29)—but there is value in providing plant support in resources inferring orthology across all domains of life. Also, plant-based analyses can benefit from the other distinctive features of OMA, such as its highly specific predictions and ability to infer HOGs. We have improved plant genome support in OMA by adding and updating more plant genomes and by inferring and annotating homeology—genes related through polyploidization—in the wheat genome.

The number of plant species in the OMA database has increased from 8 to 28 plants in recent years. In the latest release, we have added *Selaginella moellendorffii* (a lycophyte) as the deepest branching vascular plant and *Physcomitrella patens* (a bryophyte) as a representative of the non-vascular plants, thus widening the taxon set to cover ∼450 million years of plant evolution (30). We have also added the important model grass species *Brachypodium distachyon* and *Aegilops tauschii*. Additionally, we have added a variety of crop species of practical and economic importance, which are especially useful to plant geneticists and breeders. These species include: banana (*Musa acuminata subsp. malaccensis*), potato (*Solanum tuberosum*), several rice species (*Oryza brachyantha*, *Oryza glaberrima*, *Oryza sativa subsp. indica*), foxtail millet (*Setaria italica*) and bread wheat (*Triticum aestivum*).

In particular, bread wheat is the staple food source for 30% of the human population, making it one of the world's most important cereal crops. However, its very large (17 Gb), highly repetitive, hexaploid (2n = 6x = 42) genome, has made studying its organization and evolution notoriously challenging due to the lack of a high-quality reference sequence. Wheat is a recent allopolyploid resulting from two recent (<0.8 MYA ago) hybridization events between three diploid progenitors, of which the most distant pair diverged an estimated 6.5 MYA ago (31). Following that hybridization event, there has seemingly been little or no recombination across the chromosomes derived from the three progenitor genomes (32). It is therefore helpful to think of these three sets of chromosomes as 'subgenomes'. This gives rise to the notion of homeologous (also spelled 'homoeologous') chromosomes—closely related pairs of chromosomes between two subgenomes. These homeologous chromosomes have maintained a high degree of conservation amongst them, with highly similar genes located on the same chromosomal group (1 to 7) of each subgenome. However, because there have been extensive gene duplications,

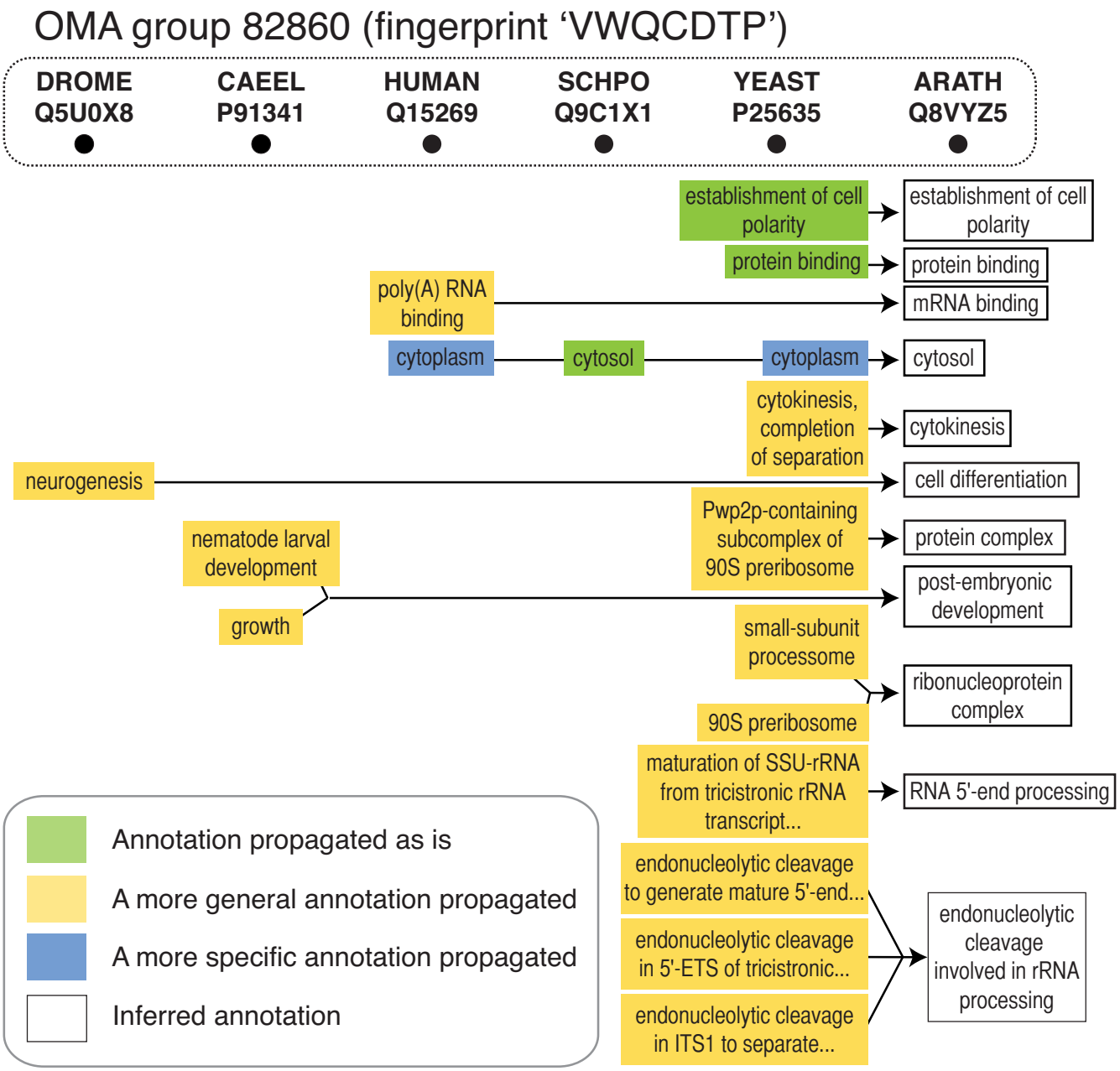


Figure 2. Gene Ontology propagation in the OMA pipeline. New Gene Ontology (GO) annotations for the sparsely annotated *Arabidopsis thaliana* protein Q8VYZ5 are inferred by propagating annotations from other members of the OMA group, taking into account implied parental terms and lineage-specific terms (see main text). For example, the inferred biological process Gene Ontology (GO) term 'post-embryonic development' is based on the more specific GO term 'nematode larval development'; the latter is in itself inappropriate to assign to a protein in the plant clade. Proteins are labelled with their SwissProt/UniProt identifiers. The abbreviations ARATH, CAEEL, SCHIPO, DROME, HUMAN and YEAST refer to species *Arabidopsis thaliana, Caenorhabditis elegans, Schizosaccharomyces pombe, Drosophila melanogaster, Homo sapiens* and *Saccharomyces cerevisiae*, respectively.

losses and rearrangements in the *Triticeae* lineage (32–35), the relationship across homeologs is not necessarily 1:1:1.

In OMA, we define homeologous genes as pairs of homologous genes that have started diverging through speciation between the progenitor genomes and then merged back into the same genome by hybridization. Thus, homeologs can be thought of as 'orthologs between subgenomes'. This suggests a simple way of adapting the OMA pipeline to infer homeologs: we first partitioned the predicted wheat proteins into the three subgenomes based on the annotation of the IWGSC (32), then inferred 'orthologs' between these subgenomes using our standard pipeline. Although conceptually straightforward, this procedure is complicated by the fragmentary nature of the current wheat survey genome, consisting of many contigs and resulting in numerous genes which are split, misannotated, or simply missing.

Dubious homeolog inferences are discarded in two steps. The first filter, part of the standard OMA algorithm, identifies instances of differential gene losses through witnesses of non-orthology in a third genome (21). This filter dis-

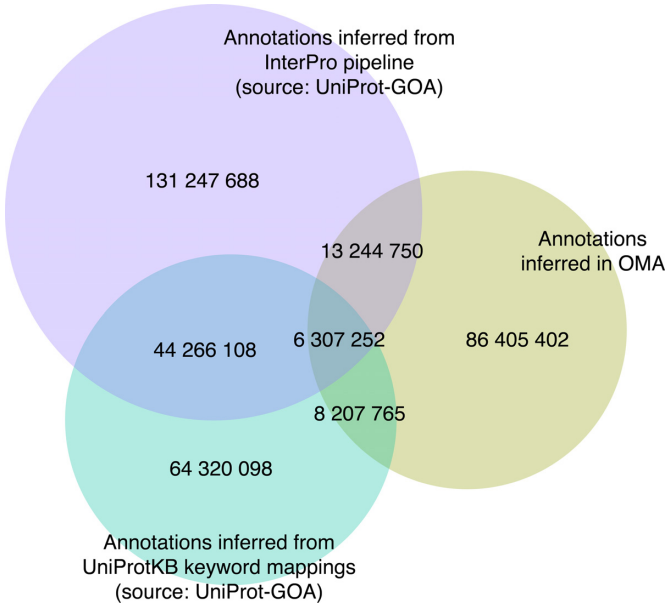


Figure 3. Numbers of electronic Gene Ontology annotations in the OMA database. Three major sources of electronic annotations are shown: annotations through the association of InterPro records with GO terms, annotations based on UniProtKB keyword mappings and annotations inferred in the OMA pipeline. The intersections show the numbers of annotations in common amongst the resources.

carded 4166 pairs. The second filter, developed specifically for homeology detection, considers the distribution of the evolutionary distances and removes outliers (defined as gene pairs with a distance higher than 2.5 standard deviations above the mean distance) from the set of reported homeologs. This discarded an additional 2212 pairs.

Two indicators suggest that the bulk of these discarded pairs are indeed unlikely to be homeologous. First, assuming that the majority of genes have remained in their ancestral position in the *Triticeae* lineage, most homeologous relationships should be between genes on corresponding chromosome groups. Yet only 14.7% of all the pairs discarded by witnesses of non-orthology and 34.7% of outliers are inferred to be between the same chromosome group (compared to 14.5% for random pairs). Second, because the three progenitor genomes diverged relatively recently (~6.5 MYA), most homeologs can be expected to be highly similar. Yet the evolutionary distance between discarded homeologous pairs is on average much higher than for the retained pairs, even if we only consider pairs filtered in the first step (Figure 4A).

We applied the same indicators to the 62 910 retained homeolog inferences. The proportion of retained homeologs involving pairs of genes on corresponding chromosome groups was considerably higher (62.8% versus 14.7–34.7% for discarded pairs). Furthermore, as expected, the distribution of evolutionary distance between predicted homeologs was skewed towards low distances, with a mean of 12.6 PAM (0.126 substitutions per site) and a standard deviation of 20.6 PAM (Figure 4B). As an additional assessment, we selected a random subset of 20 homeologous gene pairs and performed a manual validation taking into account sequence quality, gene annotation, shared chromosome group, percentage identity and evolutionary distance between pairs. Fifty-five percent of the predictions could be confirmed, with the rest being either inconclusive or likely mistakes due to misannotations (transposons, chloroplast genes), missing true homeologous counterparts, etc. (Supplementary Table S1). Given that the process of flow sorting of the wheat chromosomes and arms resulted in on average 10% contamination with other chromosomes (32), a small proportion of bona fide homeolog pairs can be expected to be erroneously annotated as belonging to different chromosome group.

In the OMA browser, retained homeolog inferences are labelled as 'high confidence' if they involve genes belonging to consistent chromosome groups, and 'low confidence' if they do not. In the latest release, this resulted in 39 442 pairs (63.2%) of high-confidence homeology predictions and 23 468 (36.8%) low-confidence ones. The average percent identity for the 12 high confidence pairs is 95.4% compared to 90.5% for low confidence pairs. We chose not to be too stringent in the cut-off for evolutionary distance and/or percent identity because although most homeolog pairs have a high degree of conservation, this might not necessarily be true for certain genes that evolve quickly such as disease resistance genes (36), transcription factors (37) or pentatricopeptide repeat proteins (38).

NEW SYNTENY VIEWER PROVIDING THE GENOMIC CONTEXT OF ORTHOLOGS

In the absence of genome rearrangement, orthology relationships can be expected to be consistent across neighbouring genes—a concept commonly referred to as 'shared synteny'. Patterns of syntenic conservation or divergence can shed light on the evolutionary history of genomic loci of interest; they can also reveal sequencing artefacts, misannotations or orthology inference errors. Synteny visualization tools have been successfully developed in several comparative genomics databases such as Yeast Gene Order Browser (39), Genomicus (40) or GnpIS (41). The OMA Browser now features a synteny viewer as well.

The OMA synteny viewer uses a typical layout: genes are represented by boxes, with neighbouring genes displayed in adjacent columns and orthologous regions displayed in different rows. The reference syntenic block, centred on a query gene, is displayed in the first row. The other rows are centred on genes that are orthologous to the query gene, ordered by increasing taxonomic distance to the query gene species. Orthology relationships to each gene contained in the reference syntenic block are coded using different colours. To convey many-to-one and many-to-many relationships, we use stripes of the relevant colours. To aid clarity, hovering over a gene highlights all orthologs of the same colour including those with stripes. The data can be conveniently explored by clicking on any gene, which recentres the display on that gene as a new query.

To illustrate the usefulness of the new synteny viewer, consider the arrangement of alcohol dehydrogenase (ADH) genes around human *ADH1A* (Figure 5). The human ADH gene cluster ADH7 (class IV)-ADH1C (class I)-ADH1B (class I)-ADH1A (class I)-ADH6 (class V)-ADH4 (class

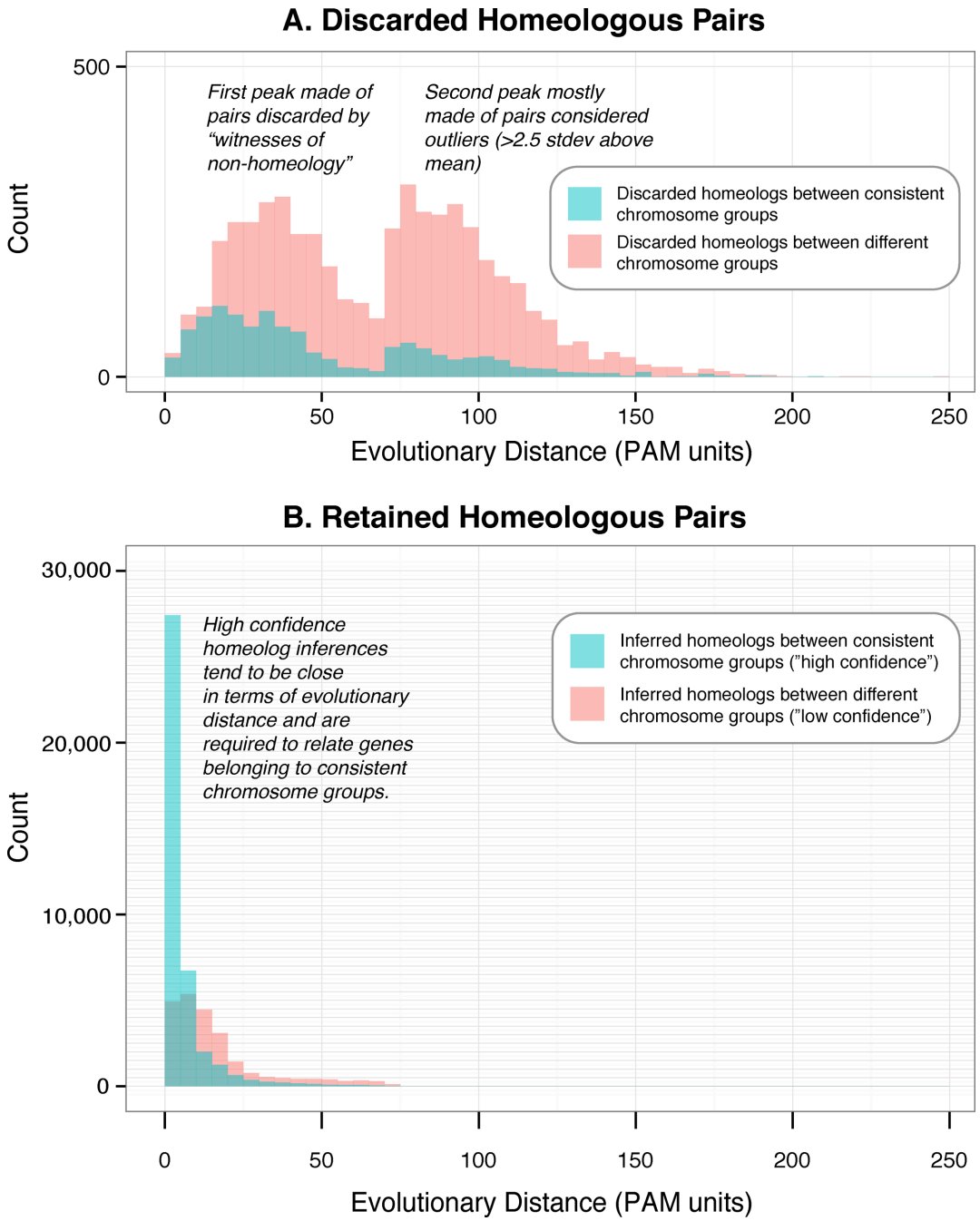


Figure 4. Distribution of evolutionary distances for homeologous pairs that were (**A**) discarded via witness of non-homeology or because they were outliers, or (**B**) retained as inferred homeologs. In both plots, the blue colour represents pairs where both homeologs are located on the same chromosome group and the red colour indicates pairs where homeologs are located on different chromosome groups. The y-axes are drawn at different scales but the grid is consistent across the two plots.

II)-ADH5 (class III) is displayed in the first row. Because the cluster sits on the complementary strand, it appears in reverse order—starting in column 3 (Gene ID 22172) and ending in column −3 (22163). The synteny viewer suggests that the neighbourhood of orthologous genes is well conserved amongst simians, but the conservation diminishes as we move to more distant lineages. Genes with stripes are in one-to-many or many-to-many orthologous relationships with human ADH1A (22168), human ADH1B (22169) and human ADH1C (22171). In particular, the presence of two orthologs in the bushbaby (OTOGA) suggests a separate duplication within the lemur lineage, yielding many-to-many orthology. These observations are all consistent with detailed analyses in the literature (42). Although positioned within well-conserved syntenic regions, genes 13367 in the chimp (PANTR) and 15069 in the gorilla (GORGO) have no human orthologous counterpart in this region. On account of their very short lengths—13 AA and 14 AA,

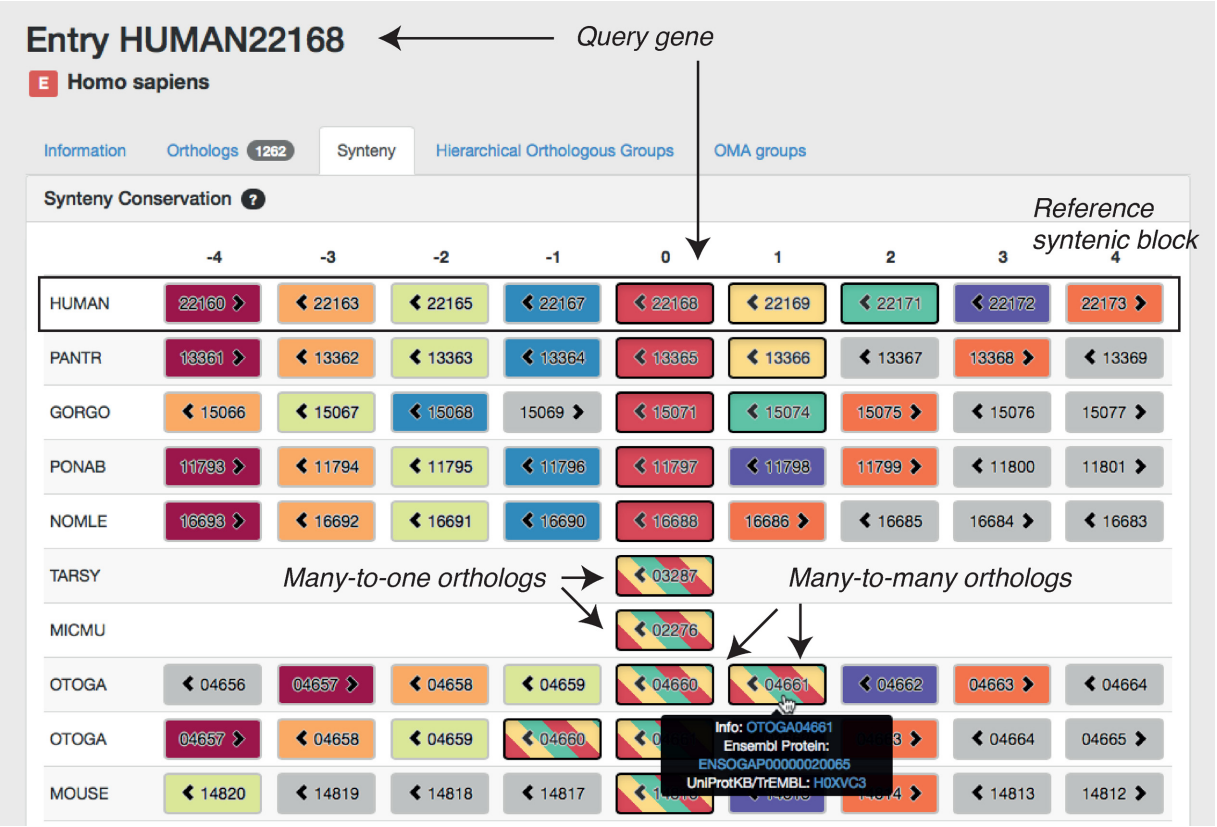


Figure 5. Screenshot of the new OMA synteny viewer with the *ADH1A* gene in human (Gene ID 22168) as query. Each gene is illustrated as a box containing a numerical OMA Gene ID and an arrow to indicate the gene's orientation. The colour of genes outside the query species indicates orthologous relationship with human genes, with bands of colour capturing many-to-one and many-to-many relationships. Genes that are non-orthologous to all nine human genes contained in this window are displayed in grey. The fragmented assemblies of tarsier (TARSY) and mouse lemur (MICMU) contain no genes next to 03287 and 02276, respectively.

respectively—they are likely to be fragments. Furthermore, the absence of flanking genes in the tarsier (TARSY) and mouse lemur (MICMU) is due to the low quality of the genome assembly in these regions.

BETTER SUPPORT FOR HOGS

As discussed above in the overview of the OMA pipeline, HOGs are a key output of the OMA algorithm; they group all the sequences that have descended from a single common ancestral gene within clades of interest. This provides an intuitive framework to generalise the concept of orthology to more than two species. For instance, if we consider the human ADH1A gene discussed in the previous section, it belongs to an HOG containing ADH1B and ADH1C as well, whilst at the more specific level of simians, the three genes belong to three distinct HOGs. This difference in resolution makes intuitive sense because as we consider a broader or narrower range of species, the shared attributes amongst them can be expected to be coarser or finer.

OMA HOGs are inferred from orthologous pairs using a fast and effective algorithm described previously (22). However, until recently, the OMA Browser had been dynamically inferring these HOGs on user demand. Large families could take a few minutes to process. Furthermore, because of the non-deterministic nature of the inference algorithm, there could be small inconsistencies for requests at different taxonomic levels (e.g. one sequence included in an HOG defined at the level of vertebrates but not included at the level of all bilateria). Starting with the latest release, HOGs are precomputed thereby providing rapid user access and consistent inferences. HOGs can now be downloaded in OrthoXML format (43) for further analyses.

One potential use of the HOGs data is to map gene losses, duplications and gains onto species trees. Indeed, since HOGs are defined in terms of ancestral genomes at all internal nodes in the species tree, keeping track of the number of HOGs and their content whilst traversing the tree can yield these quantities. Contrary to approaches solely based on gene counts in extant genomes (e.g. 44), HOGs take into account relationships between the actual sequences and thus can be expected to yield more precise estimates. Furthermore, this approach allows the user to identify the specific genes that underwent duplication or losses on particular branches of the phylogeny.

To illustrate this application, we provide an estimate of gains and losses in the primate tree obtained by parsing OMA HOGs (Figure 6). Large numbers of losses on terminal branches can be indicative of fragmentary genomes (45), such as the tarsier with its low 1.82x coverage. Even so, previous studies have reported elevated duplication and loss rates in the primate lineage (46).

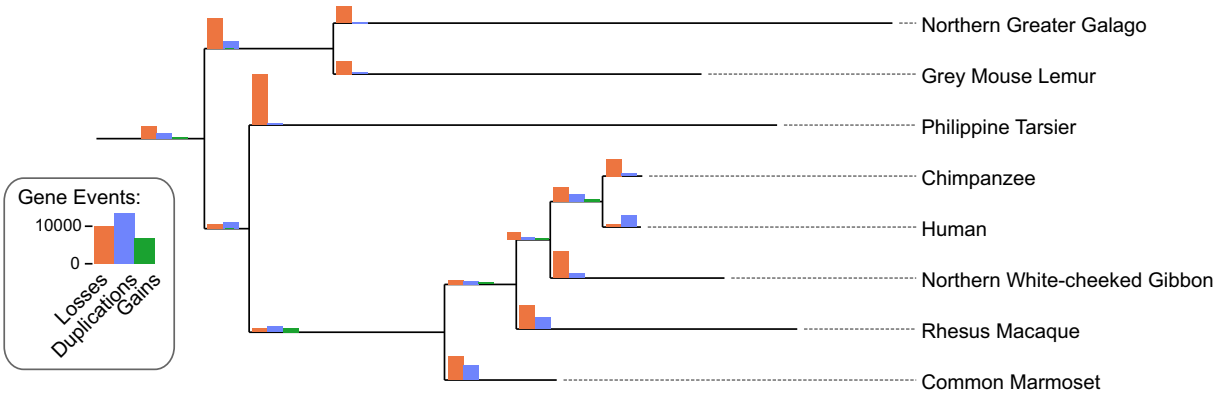


Figure 6. Gene losses, duplications and gains from hierarchical orthologous groups. Gene duplications, losses and gains on the primate lineage inferred from OMA hierarchical orthologous groups.

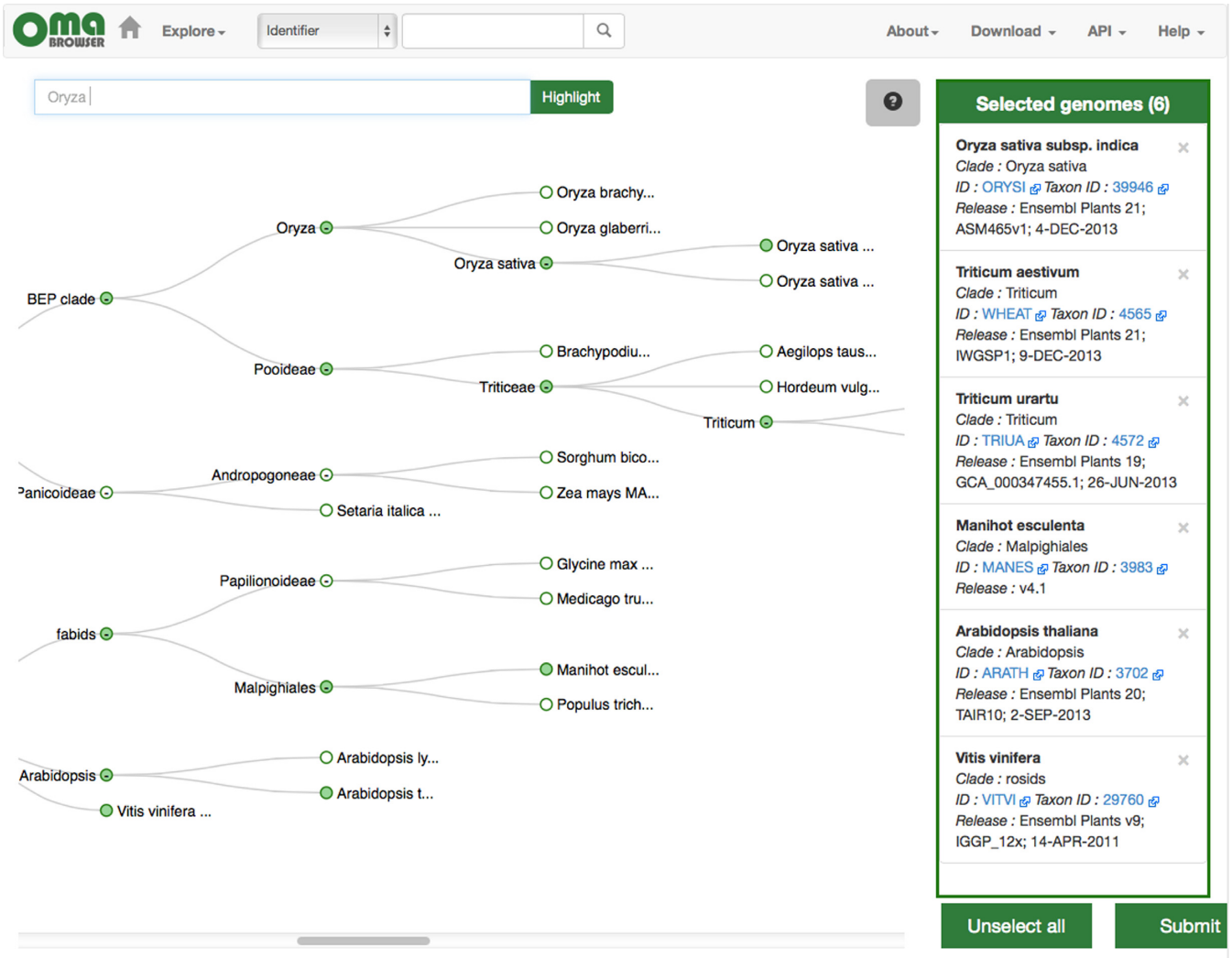


Figure 7. Selection tool for pre-computed genome export. This new function enables users to export genomes of interest and their associated all-against-all comparisons for analysis in the OMA standalone software.

EXPORT OF PROTEIN SETS AND THEIR ASSOCIATED ALL-AGAINST-ALL COMPUTATIONS

As genome and transcriptome sequencing are becoming affordable and ubiquitous, there is an increasing need for orthology prediction on custom data. As a solution to this, we have developed OMA standalone, a downloadable open source implementation of the OMA pipeline for Linux and Mac (the details of the software are the focus of a forthcoming publication). To enable users to efficiently combine custom and public genomes, we have added the possibility of exporting OMA genomes, including all-against-all computations amongst them, as input files for OMA standalone. The function is accessible via the 'Download' menu in the navigation bar of the new OMA Browser interface. Users can select up to 50 genomes for export (Figure 7), which together with OMA standalone are packaged for download as a single compressed *tar* file.

OUTLOOK

For just over a decade, the OMA database has provided orthology inference amongst complete genomes. It has remained true to its mission of providing reliable, high-quality orthology inferences across a broad taxonomic range. With 17 major releases, each including ∼100 additional and updated genomes, the project has been maintained with sustained endurance. At the same time it has also gained numerous functionalities, of which the most recent are highlighted in this update.

So what awaits OMA in the coming decade? One major challenge facing many phylogenomic resources is to keep abreast of the rapid increase in sequencing data (4). In OMA, the all-against-all protein comparison phase—the most time-consuming phase with >7 million CPU hours logged to date—grows quadratically with the number of sequences under consideration. But computational bottlenecks are nothing new in OMA; they have been a *leitmotif* all along and our experience has been that they can generally be overcome through software optimization (e.g. 47) or new heuristics (e.g. 48). We also see potential in sharing computations across different resources and have initiated a joint effort with OrthoDB (10) in that direction.

Another challenge lies with fragmentary, poorly annotated genomes and their potentially disruptive effect on orthology inference and interpretation. Yet at the same time, orthology can also help identify split genes (49). Furthermore, as discussed above, orthology combined with synteny information or integrated across multiple species in hierarchical groups can also uncover quality problems with the data.

One thing however seems certain: as the pace of genome sequencing continues to accelerate, elucidating evolutionary relationships across different genes will remain the key to exploiting the richness of this data. OMA is thus likely to stay relevant.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank three anonymous referees for their comments on the manuscript.

FUNDING

Service and Infrastructure, Swiss Institute of Bioinformatics [to G.H.G., C.D.]; UK Biotechnology and Biological Sciences Research Council [BB/L018241/1 to C.D.]; UCL Impact Award, University College London [to C.D., I.P.]; Bayer CropScience NV [to N.G., I.P.]; Biomedical Vacation Studentship, Wellcome Trust Foundation [to A.S.]; EMBL [to K.G.]. Funding for open access charge: BBSRC via the University College London Library.
Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
3. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
4. Sonnhammer,E.L.L., Gabaldón,T., Sousa da Silva,A.W., Martin,M., Robinson-Rechavi,M., Boeckmann,B., Thomas,P.D., Dessimoz,C. and the Quest for Orthologs consortium. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics*, **30**, 2993–2998.
5. Altenhoff,A.M. and Dessimoz,C. (2012) Inferring orthology and paralogy. In: Anisimova,M (ed). *Evolutionary Genomics. Methods in Molecular Biology*. Humana Press, Clifton, NJ, **855**, pp. 259–279.
6. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldón,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
7. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
8. Östlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
9. Uchiyama,I., Mihara,M., Nishide,H. and Chiba,H. (2012) MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res.*, **41**, D631–D635.
10. Waterhouse,R.M., Tegenfeldt,F., Li,J., Zdobnov,E.M. and Kriventseva,E.V. (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365.
11. Chen,F., Mackey,A.J., Stoeckert,C.J. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
12. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
13. Huerta-Cepas,J., Capella-Gutiérrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldón,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.
14. Proost,S., Van Bel,M., Sterck,L., Billiau,K., Van Parys,T., Van de Peer,Y. and Vandepoele,K. (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, **21**, 3718–3731.
15. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
16. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

17. Afrasiabi,C., Samad,B., Dineen,D., Meacham,C. and Sjölander,K. (2013) The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res.*, **41**, W242–W248.
18. Boeckmann,B., Robinson-Rechavi,M., Xenarios,I. and Dessimoz,C. (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, **12**, 423–435.
19. Linard,B., Thompson,J.D., Poch,O. and Lecompte,O. (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.
20. Roth,A.C.J., Gonnet,G.H. and Dessimoz,C. (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, **9**, 518.
21. Dessimoz,C., Boeckmann,B., Roth,A.C.J. and Gonnet,G.H. (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.*, **34**, 3309–3316.
22. Altenhoff,A.M., Gil,M., Gonnet,G.H. and Dessimoz,C. (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One*, **8**, e53786.
23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Michael Cherry,J., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
24. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Chan,W.M., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
25. Skunca,N., Bošnjak,M., Kriško,A., Panov,P., Džeroski,S., Smuc,T. and Supek,F. (2013) Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. *PLoS Comput. Biol.*, **9**, e1002852.
26. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kähäri,A. *et al.* (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
27. Monaco,M.K., Stein,J., Naithani,S., Wei,S., Dharmawardhana,P., Kumari,S., Amarasinghe,V., Youens-Clark,K., Thomason,J., Preece,J. *et al.* (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.*, **42**, D1193–D1199.
28. Rouard,M., Guignon,V., Aluome,C., Laporte,M.-A., Droc,G., Walde,C., Zmasek,C.M., Périn,C. and Conte,M.G. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **39**, D1095–D1102.
29. Van Bel,M., Proost,S., Wischnitzki,E., Movahedi,S., Scheerlinck,C., Van de Peer,Y. and Vandepoele,K. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **158**, 590–600.
30. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.-F., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
31. International Wheat Genome Sequencing Consortium, Marcussen,T., Sandve,S.R., Heier,L., Spannagl,M., Pfeifer,M., Jakobsen,K.S., Wulff,B.B.H., Steuernagel,B., Mayer,K.F.X. *et al.* (2014) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.
32. International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**, 1251788.
33. Luo,M.C., Deal,K.R., Akhunov,E.D., Akhunova,A.R., Anderson,O.D., Anderson,J.A., Blake,N., Clegg,M.T., Coleman-Derr,D., Conley,E.J. *et al.* (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl Acad. Sci. U.S.A.*, **106**, 15780–15785.
34. Akhunov,E.D., Sehgal,S., Liang,H., Wang,S., Akhunova,A.R., Kaur,G., Li,W., Forrest,K.L., See,D., Simková,H. *et al.* (2013) Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol.*, **161**, 252–265.
35. Choulet,F., Alberti,A., Theil,S., Glover,N., Barbe,V., Daron,J., Pingault,L., Sourdille,P., Couloux,A., Paux,E. *et al.* (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**, 1249721.
36. McHale,L., Tan,X., Koehl,P. and Michelmore,R.W. (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.*, **7**, 212.
37. Lagercrantz,U. and Axelsson,T. (2000) Rapid evolution of the family of CONSTANS LIKE genes in plants. *Mol. Biol. Evol.*, **17**, 1499–1507.
38. Geddy,R. and Brown,G.G. (2007) Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC Genomics*, **8**, 130.
39. Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.
40. Louis,A., Muffato,M. and Roest Crollius,H. (2013) Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res.*, **41**, D700–D705.
41. Steinbach,D., Alaux,M., Amselem,J., Choisne,N., Durand,S., Flores,R., Keliet,A.-O., Kimmel,E., Lapalu,N., Luyten,I. *et al.* (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database* , **2013**, bat058.
42. Carrigan,M.A., Uryasev,O., Davis,R.P., Zhai,L., Hurley,T.D. and Benner,S.A. (2012) The natural history of class I primate alcohol dehydrogenases includes gene duplication, gene loss, and gene conversion. *PLoS One*, **7**, e41175.
43. Schmitt,T., Messina,D.N., Schreiber,F. and Sonnhammer,E.L.L. (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.*, **12**, 485–488.
44. De Bie,T., Cristianini,N., Demuth,J.P. and Hahn,M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.
45. Milinkovitch,M.C., Helaers,R., Depiereux,E., Tzika,A.C. and Gabaldón,T. (2010) 2X genomes–depth does matter. *Genome Biol.*, **11**, R16.
46. Bailey,J.A. and Eichler,E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, **7**, 552–564.
47. Szalkowski,A., Ledergerber,C., Krähenbühl,P. and Dessimoz,C. (2008) SWPS3 - fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2. *BMC Res. Notes*, **1**, 107.
48. Wittwer,L.D., Piližota,I., Altenhoff,A.M. and Dessimoz,C. (2014) Speeding up all-against-all protein comparisons while maintaining sensitivity by considering subsequence-level homology. *PeerJ*, **2**, e607.
49. Dessimoz,C., Zoller,S., Manousaki,T., Qiu,H., Meyer,A. and Kuraku,S. (2011) Comparative genomics approach to detecting split-coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes). *Brief. Bioinform.*, **12**, 474–484.

SCIENTIFIC REPORTS

OPEN

Comparative genomics reveals contraction in olfactory receptor genes in bats

Received: 8 August 2016
Accepted: 9 February 2017
Published online: 21 March 2017

Georgia Tsagkogeorga[1], Steven Müller[2], Christophe Dessimoz[2,3,4] & Stephen J. Rossiter[1]

Gene loss and gain during genome evolution are thought to play important roles in adaptive phenotypic diversification. Among mammals, bats possess the smallest genomes and have evolved the unique abilities of powered flight and laryngeal echolocation. To investigate whether gene family evolution has contributed to the genome downsizing and phenotypic diversification in this group, we performed comparative evolutionary analyses of complete proteome data for eight bat species, including echolocating and non-echolocating forms, together with the proteomes of 12 other laurasiatherian mammals. Our analyses revealed extensive gene loss in the most recent ancestor of bats, and also of carnivores (both >1,000 genes), although this gene contraction did not appear to correlate with the reduction in genome size in bats. Comparisons of highly dynamic families suggested that expansion and contraction affected genes with similar functions (immunity, response to stimulus) in all laurasiatherian lineages. However, the magnitude and direction of these changes varied greatly among groups. In particular, our results showed contraction of the Olfactory Receptor (OR) gene repertoire in the last common ancestor of all bats, as well as that of the echolocating species studied. In contrast, non-echolocating fruit bats showed evidence of expansion in ORs, supporting a "trade-off" between sensory modalities.

Gene gain and loss are expected to be a major source of genomic variation, and a principal driver of phenotypic diversity[1,2]. Increasing evidence from whole-genome sequencing has further corroborated this hypothesis, with multiple reported cases of gene family expansion underlying evolutionary innovations in animals[3]. Concurrently, large-scale population genomic studies in humans have shown that gene copy number variation can result in a range of pathologies or diseases[4], highlighting further the contribution of changes in gene family size on phenotype[5].

Mammalian genome evolution is characterized by multiple episodes of expansion and contraction[6]. A mammalian genome contains an average of around 3.14 Gb, however, genome size can reach 6.18 Gb in some members of the Afrotheria[7]. Of all mammals, bats possess the smallest genomes, with an average recorded content of 2.36 Gb (± 0.28 Gb S.E.), and a recorded minimum size of ~1.59 Gb (or 1.63 pg) in Carriker's round-eared bat *Lophostoma carrikeri*[8]. It was previously shown that genome size reduction in bats can be partially attributed to shortened introns and intergenic regions, a trend that is also seen in birds[9]. In both groups, it has been suggested that genome contraction might be an adaptation for powered flight and its associated high metabolic rates[10].

In terms of the actual gene content, insights into the dynamics of gene gain and loss come mainly from genomic studies based on large-scale sequencing projects. Comparative analysis of the genomic sequences of the mouse-eared bat, *Myotis davidii*, and black flying fox, *Pteropus alecto*, with a range of other mammals, identified episodes of gene expansion in 71 gene families in *M. davidii* and 13 in *P. alecto*, as opposed to 41 and 35 contractions, respectively[11]. Likewise, it has been reported that 44 families have experienced gene loss, and 67 gene expansion, in the genome of the Brandt's bat *Myotis brandtii*[12].

Despite the large amount of genomic data generated for bats over the past four years[11–13], a comprehensive picture of gene gain and loss at a genome scale for the group is still lacking. This is mainly because all studies so far have focussed on one or two representative species at a time, or, when encompassing more taxa, have been restricted to specific gene families[14–16]. Here we investigate the patterns of gene family evolution in bats in greater

[1]School of Biological & Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK. [2]University College London, Gower Street, London, WC1E 6BT, UK. [3]University of Lausanne, Biophore, 1015, Lausanne, CH, Switzerland. [4]Swiss Institute of Bioinformatics, Biophore, 1015, Lausanne, CH, Switzerland. Correspondence and requests for materials should be addressed to G.T. (email: g.tsagkogeorga@qmul.ac.uk) or S.J.R. (email: s.j.rossiter@qmul.ac.uk)

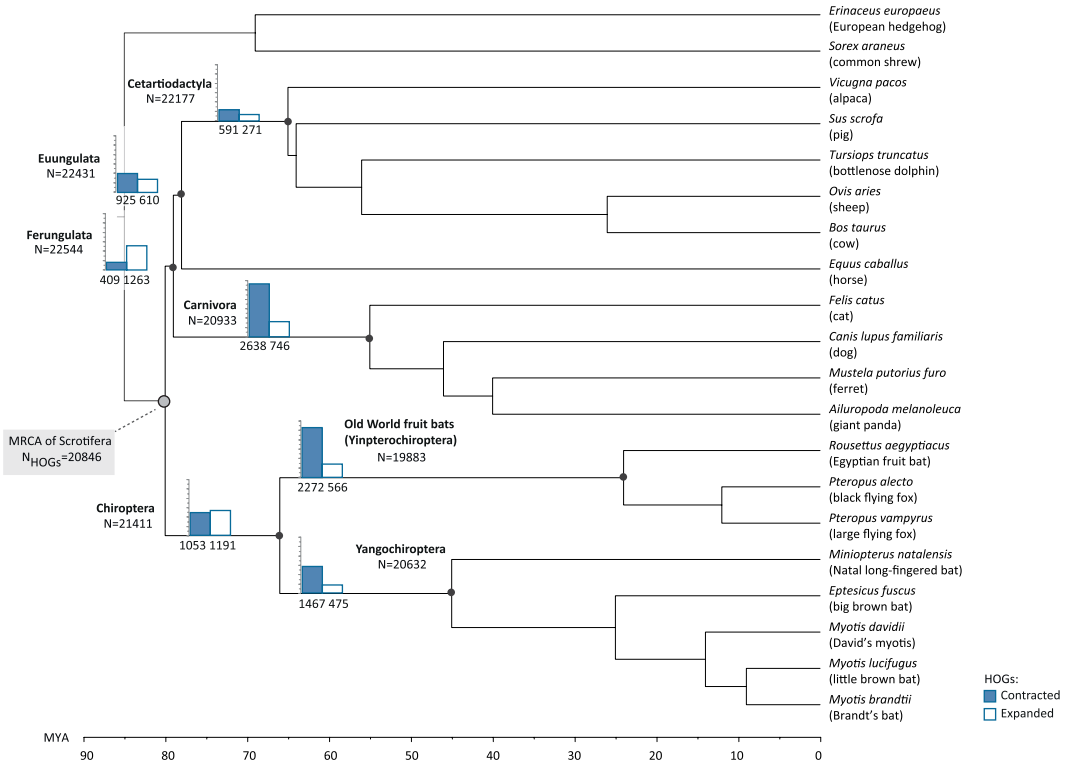


Figure 1. Gene losses and gains along key branches of the Laurasiatheria phylogeny, as inferred from OMA hierarchical groups.

depth. Using a comparative genomics approach spanning 20 mammalian genomes, we assess the average turnover of gene gain and loss in bats, and test whether this rate is different from that of other closely related lineages. We also aim to identify families that have undergone accelerated evolution in the last common ancestor of bats, as well as in the last common ancestor of echolocating and non-echolocating forms. Finally, we address the question of whether rates of change in gene family size in bats are associated with their unusually small genome sizes.

Results and Discussion

Taxon sampling and inference of Hierarchical Orthologous Groups (HOGs) across mammals. To identify homologous genes and elucidate evolutionary patterns of gene gain and loss in bats, we compared the complete proteomes from 20 laurasiatherian mammals using the Orthologous MAtrix (OMA) algorithm implemented in the OMA standalone software package[17]. Our sampling included eight bats, of which three are non-echolocating Old-World fruit bats from the suborder Yinpterochiroptera (*Pteropus alecto*, *P. vampyrus* and *Rousettus aegyptiacus*), and five are echolocating species from the suborder Yangochiroptera (*Miniopterus natalensis*, *Eptesicus fuscus*, *Myotis brandtii*, *M. davidii* and *M. lucifugus*). To assess completeness of the laurasiatherian proteomes, we used the software BUSCO[18]. The estimated completeness across our sampled species gene annotations varied from 55% for *Sorex araneus* to 98% for *P. alecto*. All bat proteomes were 90% complete, with levels of fragmentation varying from 1% for *P. alecto* and *R. aegyptiacus* to 4.9% for *M. davidii* (Supplementary Table S1).

For our OMA analysis, we used the species tree topology based on the most recent phylogenomic study of bats[13], in which bats were found to be most closely related to the Ferungulata, i.e., the clade uniting carnivores and ungulates (Fig. 1). However, the phylogenetic relationship among laurasiatherian lineages remains highly contentious, attributed largely to their rapid radiation ~80 million years ago and associated extensive incomplete lineage sorting[13]. As a result, many competing phylogenetic scenarios have been proposed for the diversification of laurasiatherian mammals, and thereby the phylogenetic positioning of bats within them[19–25]. To account for the effects of species phylogeny on inferring gene gains and losses in bat genomes, we also repeated our analyses under six proposed alternative species tree topologies for Laurasiatheria (Supplementary Figure S1), drawn from the recent literature[19–25].

Using all-against-all similarity searches coupled with a graph-based clustering approach, OMA grouped the 20 mammalian proteomes into 20,936 Hierarchical Orthologous Groups (HOGs—sets of all genes descended from a common ancestral gene within the Laurasiatheria). We inferred 32,571 orthologous genes across all sampled species, including 26,537 with at least one bat sequence.

Gene gain and loss in bats and other major laurasiatherian lineages. Based on HOG information, we mapped gene duplications and losses on our mammalian tree (Fig. 1). Considering key branches of our phylogeny, our analyses identified 14,445 HOGs present in most recent common ancestor (MRCA) of all

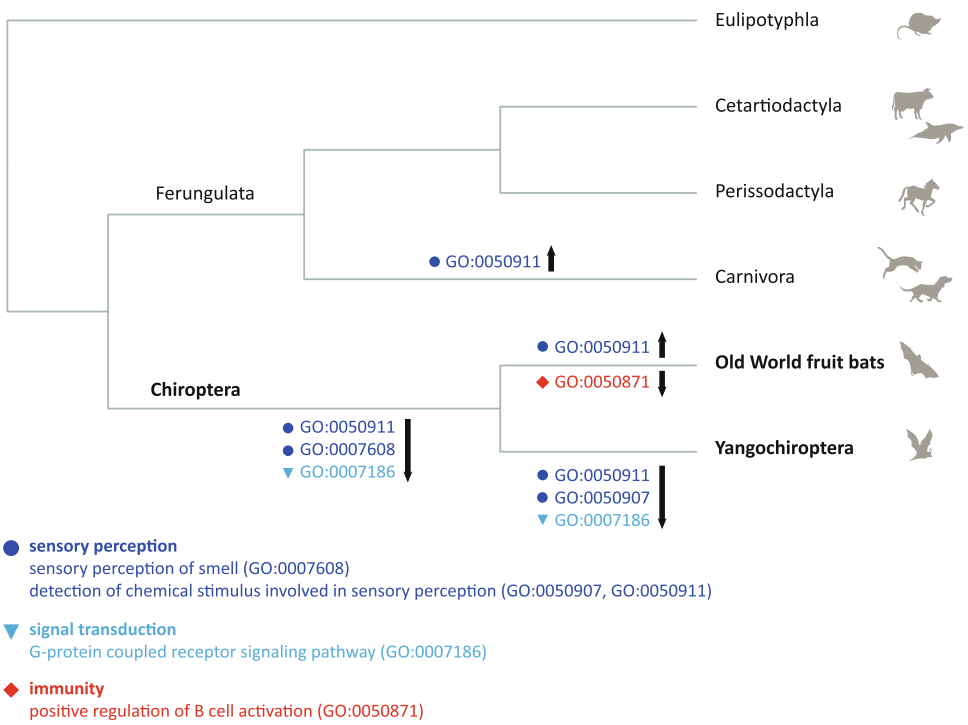


Figure 2. Significant GO terms associated with expanded and contracted HOGs in bats and other laurasiatherians (Ontology: Biological Process; Bonferroni corrected p-value < 0.05). Arrows indicate expansion (upward) and contraction (downward).

sampled mammals, and 20,846 HOGs in the branch leading to the MRCA of the clade uniting bats, carnivores and ungulates (clade of Scrotifera). We obtained 22,544 HOGs in the MRCA of carnivores and ungulates (clade of Ferungulata) and 21,411 HOGs in the MRCA of bats (Chiroptera) (Fig. 1).

Inference of ancestral gene duplication events across Laurasiatheria revealed 1,191 episodes of ancestral gene expansion in the MRCA of bats, resulting in 2,678 multi-copy genes since their divergence from carnivores and ungulates (MRCA of Scrotifera in Fig. 1). Very similar levels of gene duplication were also inferred for the latter two groups, with 1,263 expansions predicted at the level of the MRCA of Ferungulata (Fig. 1; Supplementary Table S2).

In terms of gene loss, the MCRA of bats showed evidence of having undergone more extensive loss compared to that of the Ferungulata, with 1,191 gene losses compared to 409, respectively. Within the Ferungulata, high numbers of HOG contractions were also inferred along the branch leading to the MRCA of carnivores (n = 2,638). The fewest changes in gene family size were observed along the branch of the MRCA of Cetartiodactyla, with 591 HOGs showing evidence of gene loss, nearly half of that predicted in bats (Fig. 1; Supplementary Table S2).

Functional profiling of expanded and contracted gene families in bats. To examine the biological role of HOGs that show changes in number in bats, we first assigned one representative member of each HOG to GO terms in UniProtKB[26] (Supplementary Tables S3–S5). Among HOGs that appeared to have expanded, the vast majority were categorised as being genes involved in biological regulation (>400 HOGs), metabolic genes (>300 HOGs), and genes associated with a response to stimulus (>200 HOGs). We also obtained a strong signal of expansion for genes involved in developmental process (>100 HOGs), cellular component organization (>100 HOGs), immunity (80 HOGs), locomotion (>40 HOGs), and reproduction (>40 HOGs) (Supplementary Table S4). Similar functional profiles were obtained for HOGs showing evidence of gene loss in bats, with again a substantial number of gene clusters linked to metabolic process (>300 HOGs) and response to stimulus (~200 HOGs), as well as cellular localisation (>100 HOGs), developmental process (>100 HOGs) and immune system (>50 HOGs) (Supplementary Table S4).

To gain further insights into putative roles of genes duplicated and lost in bat genomes, we looked at the annotation of their homologues in cow and human genomes (Supplementary Table S3) and carried out GO enrichment analyses to test for overrepresented functional terms associated with expanded and contracted HOGs in bats (Supplementary Tables S6 and S7). We did not find any significantly enriched GO terms for expanded genes in the MRCA of bats after correction for multiple testing (Supplementary Table S6), however, our analyses indicated that genes encoding proteins with olfactory receptor activity (GO:0004984), including those involved in the sensory perception of smell (GO:0050911), were significantly enriched among contracted HOGs in bat genomes (Fig. 2, Supplementary Table S7). This reduction in numbers of OR genes was corroborated using both the cow (uncorrected p-value = 3.26E-10; p-value Bonferroni corrected = 4.34E-06) and human homologues (uncorrected p-value = 8.70E-11; p-value Bonferroni corrected = 1.42E-06). GO results based on human also showed

significant enrichment (p-value Bonferroni corrected <1E-05) in genes associated with G-protein coupled receptor activity (GO:0004930) and signaling pathway (GO:0007186), sensory perception of smell (GO:0007608), as well as in genes with products categorised as odorant binding (GO:0005549 in Supplementary Table S7).

Of all 1,053 HOGs inferred as being reduced in size in bats, we identified 357 HOGs that appeared to be completely absent in bats (i.e. with no homologous sequences detected). Of these, a few consisted of ambiguous groupings comprising two to three (usually partial) sequences from other mammal genomes that presumably failed to cluster together with the remaining HOGs. To filter genuine signal of gene loss from potential clustering artifacts, we identified HOGs that contained no bat genes yet included a sequence from at least three of the laurasiatherian lineages sampled (i.e. from Carnivora, Cetartiodactyla, Perissodactyla and Eulipotyphla). We identified 201 such cases in total, with most having functional links with biological regulation (92 HOGs), response to stimulus (83 HOGs), and signal transduction (76 HOGs). We also performed clustering of these HOGs using a keyword clustering approach from protein description available in UniProtKB, which recovered associations for 73 HOGs to biological processes of olfaction (68), transcriptional regulation (3), differentiation (1), myogenesis (1) transport (1) and Ubl conjugation pathway (1). Looking at the protein identifiers *per se*, we confirmed that the vast majority of HOGs with no orthologues in bats consisted of Olfactory Receptor (OR) genes (Supplementary Table S5).

For any given lineage, identifying genes that are novel is inherently more difficult than identifying ones that are either lost or duplicated. This is because novel genes are usually not associated with any functional annotation, and also because they might be present in unsampled taxa. Thus to characterise putative novel genes present only in bats, we manually inspected representative gene identifiers for corresponding HOGs and functionally classified these loci either on the basis of similarity to other known genes (e.g. "MHC class II transactivator-like" gene) or the presence of common protein motifs (e.g. zinc finger protein like). We identified 131 HOGs containing putatively novel genes, 37 corresponded to low quality or/and uncharacterised proteins. Of the remaining HOGs, at least 11 had a binding activity, six corresponded to transport proteins, five had functional links to immunity, and two to sensory perception of olfaction. We further identified HOGs with putative novel genes encoding common protein domains, such as zinc finger (6), growth factors (2) and IQ protein domains (3).

Functional annotation of HOG size changes in other lineages. We next assessed whether the patterns of HOG expansion and contraction in bats with respect to gene types were also seen in other laurasiatherian groups. For this, we conducted GO enrichment analyses of HOGs showing expansion and contraction in the MRCAs of Ferungulata, and its constituent lineages Cetartiodactyla and Carnivora (Supplementary Tables S8–S13). We found no significant GO terms after correcting for multiple tests in the MRCAs of Ferungulata and cetartiodactyl mammals. In contrast, however, genes encoding proteins with olfactory receptor activity (GO:0004984) associated with the sensory perception of smell (GO:0050911) were enriched among HOGs showing expansion in the MRCA of carnivores (uncorrected p-value = 8.38E-09; p-value Bonferroni corrected = 0.0001); this trend in carnivores was thus opposite to that observed in bats (Fig. 2, Supplementary Table S10). In the same branch, our results suggested expansion also in genes encoding for ribosomal proteins (uncorrected p-value = 1.50E-06; p-value Bonferroni corrected = 0.0244). Some evidence for expansion in OR genes was also detected for the MRCAs of Ferungulata and Cetartiodactyla although with low statistical support (uncorrected p-value = 0.0004 in Supplementary Table S8, and uncorrected p-value = 0.0080 in Supplementary Table S12). We found no evidence of enrichment associated with sensory perception genes for contracted HOGs along the ancestral branches of Ferungulata, Carnivora or Cetartiodactyla (Supplementary Tables S9, S11 and S13).

HOG evolution in echolocating and non-echolocating bats. Within bats, our analyses inferred 19,883 HOGs in the MRCA of the three non-echolocating Old-World fruit bats, *R. aegyptiacus*, *P. alecto* and *P. vampyrus*, and 20,632 in the MRCA of the clade uniting the echolocating bats *M. natalensis*, *E. fuscus*, *M. brandtii*, *M. davidii* and *M. lucifugus*, all of which are members of the suborder Yangochiroptera (Fig. 1).

In terms of gene expansion, we inferred 1,283 and 1,076 multi-copy genes in the ancestral branches of Old World fruit bats and Yangochiroptera, respectively, which were clustered into 566 and 475 respective HOGs (Fig. 1; Supplementary Table S2). Functional annotation of expanded HOGs using UniProtKB suggested an increase in genes involved in biological regulation, metabolic processes, response to stimulus, development, localization, and immunity in both echolocating and non-echolocating bat genomes (Supplementary Table S14).

GO enrichment analyses based on human homologues showed a significant expansion in OR genes (GO:0050911, GO:0004984) in the MRCA of Old World fruit bats (uncorrected p-value = 3.99E-11; p-value Bonferroni corrected = 6.50E-07), which was not found in the MRCA of the echolocating species examined (Fig. 2, Supplementary Tables S15 and S16). Similar results, were also observed when annotation was based on cow homologues, together with expansion in genes involved in respiratory system process (GO:0003016) and the regulation of defense response to viruses (GO:0050691), although all four terms showed low statistical support after correction (Supplementary Table S15). Top GO terms for HOGs expanded in Yangochiroptera genomes suggested expansion in genes linked to cell fate determination (GO:0007493, GO:0042074), lipid digestion (GO:0044241) and the regulation of immune response (GO:0050776), although again with low statistical support (Supplementary Table S16).

Compared to gene gain (n = 566), we detected a four-fold increase in gene loss in Old World fruit bats (n = 2,272) (Fig. 1; Supplementary Table S2). HOGs showing contraction along the MRCA of Old World fruit bats were associated with 50 GO terms based on human homologues (uncorrected p-value < 0.01, Supplementary Table S15). Of these, top ranked gene clusters had a direct link with immunity and pathogen recognition (GO:0050871, GO:0003823, GO:0006958, GO:0034987, GO:0042571, GO:0006910, GO:0059776,). In particular, our analyses suggested a significant enrichment of genes involved in B cell activation (GO:0050871, uncorrected p-value = 1.00E-06; p-value Bonferroni corrected = 0.0163) and antigen binding (GO:0003823, uncorrected

p-value = 2.88E-06; p-value Bonferroni corrected = 0.0468) among HOGs showing contraction along the branch of the MRCA of Old World fruit bats (Fig. 2, Supplementary Table S17). Other contracted HOGs in this group were associated with metabolism (e.g. GO:0006706, GO:2001303). Functional annotation of contacted HOGs based on cow also suggested a potential contraction in genes involved in sensory perception of smell (GO:0007608), as well as genes associated with development and morphogenesis (e.g. GO:0061326, GO:0072144, GO:0003278, see Supplementary Table S18).

Echolocating bats from the suborder Yangochiroptera also showed greater gene loss (n = 1,467) than gain, although this was less pronounced than in the fruit bats (Fig. 1; Supplementary Table S2). GO analyses of gene losses along the ancestral branch of these bats firmly supported enrichment in genes involved in detection of chemical stimulus and sensory perception of smell (GO:0050911, p-value Bonferroni corrected = 9.57E-07), and for OR genes (GO:0004984, p-value Bonferroni corrected = 9.57E-07) in particular (Fig. 2, Supplementary Table S18). Robust statistical support was also obtained for contraction in the G-protein coupled receptor signaling pathway (GO:0004930, GO:0007186, p-value Bonferroni corrected <1E-08 in Supplementary Table S18).

Overall, while Old World fruit bats showed a much greater extent of gene contraction than did echolocating species, we cannot rule out the possibility that this difference might reflect differences in the sampling size of the two groups in our analyses (three Old World fruit bats sampled versus five yangochiropterans). Despite this, the functional profiles of genes lost in these two groups are distinctive from each other, and it is less plausible that these are driven solely by potential biases in the data (Supplementary Tables S17 and S18).

With regards to novel gene gain, we identified 27 putatively novel genes in the MRCA of Old World fruit bats. Among these, eight seem to encode for binding proteins and four are associated with a response to stimulus (two involved in defence and two in olfaction) and three with metabolism. In the MRCA of Yangochiroptera, our analyses inferred 87 putative gene gains, including functions related to immune system (~30), metabolism (~40) and responses to stimulus (~50).

Effects of species phylogeny on gene expansions and contractions in bats. Our results showed that estimates of HOG expansions and contractions along the MRCA of bats varied widely among different tree topologies, with the number of gene expansions ranging from 350 to 1214 HOGs and gene losses from 768 to 1861 depending on the topology used (Supplementary Table S19). On the contrary, estimates of gene gain and loss in the MRCA of Old World fruit bats and that Yangochiroptera were relatively unaffected by the species tree, showing almost no variation across the six topologies tested (Supplementary Figure S1, Supplementary Table S19).

To assess further the robustness of our results across alternative topologies, we repeated the GO analyses for the two trees that showed the greatest differences in estimates of gene gain and loss, both compared to our results, and to each other (Trees #2 and #3 in Supplementary Figure S1 and Supplementary Table S19). These results also firmly supported a contraction in OR genes (GO:0004984, p-value Bonferroni corrected <2.3E-06) and in genes involved in sensory perception of smell (GO:0050907, GO:0050911, p-value Bonferroni corrected <3.5E-05) in both the MRCA of bats and that of echolocating taxa (Supplementary Table S20). Moreover functional annotation of HOGs corroborated both the observed expansion of olfactory genes in Old World fruit bats and Carnivores (GO:0004984, GO:0050911, corrected p-values < 4.8E-04) (Supplementary Table S21). Finally, under the two alternative trees we found a stronger signal for changes in immunity genes for across all lineages tested (Supplementary Tables S20 and S21), in line with expectations that immunity-related gene families show dynamic patterns of evolution in mammals.

Maximum likelihood estimation of genomic turnover. Although the mechanism by which multiple gene copies are generated and maintained is debated, it has been shown that many gene families follow a birth-and-death mode of evolution, including immune gene families and sensory receptor superfamilies[27,28] According to this model, genes are created by gene duplication and some are maintained in the genome for a long time, whereas others are deleted or become nonfunctional through deleterious mutations[29]. In addition to this, draft genome assemblies are prone to extensive errors in predictions of number of genes due to genome fragmentation and thus poor quality gene annotations[30], which may result in either an overestimation or underestimation of the gene load. An overestimation in the number of genes present in such genomes may be caused by the splitting of alleles into separate scaffolds or contigs. Conversely, the opposite may also happen whereby copy number variants or recent paralogues are erroneously collapsed together into a single genomic region, leading to an underestimation of gene copy number.

To account for potential erroneous estimates of the numbers of inferred homologous groups arising from the types of errors described above, we fitted to our data a recently developed model of gene family evolution that allows for estimation and correction of annotation errors from incomplete genomes[31]. We calculated the global error in our HOG dataset, as well as error for each species separately. Global error estimation in the 20,936 HOGs was found to be 0.0635, suggesting an average of 6.35% error in our HOG size measures at the tips of our tree. However, error estimates varied widely among sampled genomes, with individual species error ranging from 0 (e.g. *M. natalensis*) to 0.4126 (e.g. *Vicugna pacos*). Aside from the *M. natalensis* data, error for the remaning bats spanned from 0.0130 for *E. fuscus and R. aegyctiacus* to 0.1396 for the *M. lucifugus* genome.

To assess the average rate of gene gain (λ) and loss (μ) at a genome-scale and across our entire tree, the bat clade, as well as in the Old World fruit bats and Yangochiroptera, we analysed a subset of 18,698 HOGs that had at least one representative member present in the MRCA of Scrotifera (bats + carnivores + perissodactyls + cetartiodactyls) in a ML framework. Using birth-and-death models, we first estimated the expected number of changes in HOG size across our tree. Assuming equal rates of gene gain and loss, and accounting for errors at an individual species level, the average rate of gene turnover across our laurasiatherian phylogeny was estimated to be 0.0008 changes/gene/million years (Mya) (Table 1, one-lambdamu model + individual species error correction).

BD Model	λ_0	λ_1	λ_2	μ_0	μ_1	μ_2	-lnL	#HOGs $p^* < 0.05$	#HOGs $p^* < 0.01$	#HOGs $p^* < 5.35e\text{-}7$
No error correction										
1-lambda	$\lambda_0 = 0.0017$	—	—	—	—	—	184,044.80	3,032	1,937	188
2-lambda	$\lambda_0 = 0.0017$	$\lambda_{bats} = 0.0018$	—	—	—	—	184,028.60	3,129	1,923	518
3-lambda	$\lambda_0 = 0.0016$	$\lambda_{yan} = 0.0021$	$\lambda_{owf} = 0.0017$	—	—	—	183,868.78	2,904	1,781	1,010
1-lamdamu	$\lambda_0 = 0.0004$	—	—	$\mu_0 = 0.0030$	—	—	167,963.90	3,357	1,945	851
2-lamdamu	$\lambda_0 = 0.0003$	$\lambda_{bats} = 0.0004$	—	$\mu_0 = 0.0029$	$\mu_{bara} = 0.0031$	—	167,925.19	3,352	2,368	875
3-lamdamu	$\lambda_0 = 0.0008$	$\lambda_{yan} = 0.0007$	$\lambda_{owf} = 0.0005$	$\mu_0 = 0.0029$	$\mu_{yan} = 0.0035$	$\mu_{owf} = 0.0030$	167,644.44	3,444	2,118	614
Global error correction, $\varepsilon = 0.0635$										
1-lambda	$\lambda_0 = 0.0008$	—	—	—	—	—	170,416.03	3,126	2,200	1,285
2-lambda	$\lambda_0 = 0.0010$	$\lambda_{bats} = 0.0006$	—	—	—	—	170,086.71	4,673	3,592	2,817
3-lambda	$\lambda_0 = 0.0011$	$\lambda_{yan} = 0.0004$	$\lambda_{yan} = 0.0007$	—	—	—	169,283.74	4,995	3,924	2,834
1-lamdamu	$\lambda_0 = 0.0001$	—	—	$\mu_0 = 0.0019$	—	—	160,043.58	5,346	4,182	2,421
2-lamdamu	$\lambda_0 = 0.0001$	$\lambda_{bats} = 0.0001$	—	$\mu_0 = 0.0021$	$\mu_{bara} = 0.0014$	—	159,679.42	5,304	3,920	2,732
3-lamdamu	$\lambda_0 = 0.0001$	$\lambda_{yan} = 0.0001$	$\lambda_{owf} = 2.24e\text{-}05$	$\mu_0 = 0.0023$	$\mu_{yan} = 0.0010$	$\mu_{owf} = 0.0008$	159,082.23	5,597	4,429	2,799
Individual species error correction, $\varepsilon = [0, 0.4126]$										
1-lambda	$\lambda_0 = 0.0008$	—	—	—	—	—	162,987.48	5,076	3,980	2,336
2-lambda	$\lambda_0 = 0.0008$	$\lambda_{bats} = 0.0008$	—	—	—	—	162,987.36	5,144	4,259	2,921
3-lambda	$\lambda_0 = 0.0008$	$\lambda_{yan} = 0.0007$	$\lambda_{owf} = 0.0005$	—	—	—	162,836.70	5,006	3,682	2,746
1-lamdamu	$\lambda_0 = 0.0001$	—	—	$\mu_0 = 0.0016$	—	—	155,133.69	5,431	3,806	1,503
2-lamdamu	$\lambda_0 = 0.0001$	$\lambda_{bats} = 0.0001$	—	$\mu_0 = 0.0017$	$\mu_{bara} = 0.0015$	—	155,107.15	5,192	4,093	2,699
3-lamdamu	$\lambda_0 = 0.0001$	$\lambda_{yan} = 0.0002$	$\lambda_{owf} = 0.0001$	$\mu_0 = 0.0018$	$\mu_{yan} = 0.0013$	$\mu_{owf} = 0.0011$	154,901.51	5,499	4,517	2,547

Table 1. Overview of birth-and-death (BD) models fitted on 18,698 HOGs present in the MRCA of Scrotifera (Bats + Carnivores + Perissodactyla + Cetartiodactyla) estimated using the OMA pipeline. *Describes the likelihood of the observed sizes given the rates of gain and loss.

Breaking this down further, we obtained estimates of the global rates of gene duplication ($\lambda = 0.0001$ duplications/gene/Mya) and loss ($\mu = 0.0016$ losses/gene/Mya) (see Table 1, one-lambdamu model + individual species error correction).

We confirmed that the estimate of gene turnover was much higher when error in the data was not taken into consideration in the ML analyses (one-lambda model with no correction: $\lambda = \mu = 0.0017$) or when only global error was accommodated (one-lambda model + global error correction: $\lambda = \mu = 0.0008$). The model that incorporated error at a species-level gave the best fit to our data based on the Akaike Information Criterion (AIC) (Supplementary Table S22).

We also sought to estimate the average rate of gene turnover for the bat clade using two-lambda models. Our results suggested an equal average genomic turnover for bats compared to the rest of the Laurasiatheria after accommodating potential error ($\lambda_{bats} = \mu_{bats} = 0.0008$). Although the observed turnover for bats and the rest of laurasiatherian mammals was estimated to have the same value, a model comparison based on simulations (Supplementary Figure S2) and the AIC supported a higher probability of a gene being lost in the other laurasiatherians than in bats (best model two-lambdamu: $\lambda_{bats} = 0.0001$, $\lambda_0 = 0.0001$; $\mu_{bats} = 0.0015$, $\mu_0 = 0.0017$; ΔLnl = 26.54 [critical value 3.627, $p = 0.05$] in Supplementary Figure S2, lowest AIC in Supplementary Table S22). A similarly lower rate of gene loss was also detected in our control analyses for carnivores ($\lambda_{carnivores} = 0.0001$, $\lambda_0 = 0.0001$; $\mu_{carnivores} = 0.0014$, $\mu_0 = 0.0017$), whereas cetartiodactyls showed the inverse relationship, *i.e.*, an increased rate of average gene loss (two-lambdamu: $\lambda_{cetartiodactyls} = 0.0001$, $\lambda_0 = 0.0001$; $\mu_{cetartiodactyls} = 0.0019$, $\mu_0 = 0.0016$, in Supplementary Table S22).

Finally, we looked at the rates of change in HOG size within bats by fitting a three-lambda model, in which we specified different rates of gene family turnover for non-echolocating Old World fruit bats (λ_{owf}), for echolocating forms from the suborder Yangochiroptera (λ_{yan}), and for the rest of the tree (λ_0). With and without error correction, our results suggested a large difference in the amount of gene gain and loss between echolocating and non-echolocating forms, with genomic turnover being higher in the former (suborder Yangochiroptera $\lambda_{yan} = 0.0007$) compared to the latter (Old World fruit bats $\lambda_{owf} = 0.0005$). This trend held when rates of gene expansion were separately estimated from rates of contraction (best-fit model: 3-lamdamu + species error correction in Table 1), with estimates of both λ_{yan} and μ_{yan} being higher that corresponding values of λ_{owf} and μ_{owf}.

To assess whether the observed gene turnover in bats could help to explain their small genome sizes, we calculated the average gene expansion and expected number of gene gains and losses for each of the eighteen terminal branches of our phylogeny under the one-lambda model with error correction at a species-level. Correlating these estimates with respective genome sizes revealed no significant correlation, either before or after accounting for phylogenetic affiliation. Thus we found no evidence that broad trends in gene loss and gain have contributed to genome contraction in bats, at least for the set of taxa and the 18,698 groups of homologues studied.

Accelerated evolution of HOGs in echolocating and non-echolocating bats. We calculated the probability for each HOG evolving under the stochastic birth-and-death process, the λ rate, as well as the mean

number of gene gain or loss per HOG along each branch of the species tree (data not shown). Among 18,698 HOGs examined, we found that 2,336 were highly unlikely to have evolved under a random gain and loss process (corrected p-value < 0.01), instead showing evidence of accelerated evolution. Of these non-randomly evolving HOGs, we identified 533 showing rapid evolution along at least one bat branch (branch-specific p-value < 0.01), 130 in the Old World fruit bat clade and 460 in the clade of Yangochiroptera. Functional annotation of these 533 HOGs revealed putative associations with signal transduction (108), in particular in G-protein coupled receptors (75) such as OR and taste receptors. Evidence of accelerated evolution was also found in genes encoding proteins of the immune system (22), including endogenous retroviral elements, interferon, MHC, T-cell receptor families, and gene families involved in organ development (21) and reproduction (10).

For comparison, we examined non-randomly evolving HOGs in carnivores and cetartiodactyl mammals. We identified 248 HOGs showing accelerated evolution in the Carnivora and 537 in the Cetartiodactyla (branch-specific p-value < 0.01). Functional annotation of these groups revealed mainly metabolic and OR types of genes in carnivores, in line with our results of HOG expansion obtained from the OMA pipeline. Similar to bats, accelerated rates of gene turnover were also detected in immunity related genes and reproductive proteins in both Carnivora and Cetartiodactyla.

Gene turnover and insights into bat biology. Our analyses of HOGs found that the most highly dynamic gene families in terms of turnover across all laurasiatherian lineages were related to olfaction. In particular, we found a strong signature of contraction of OR genes in all bats (Fig. 2, Supplementary Table S7), which was also associated with an accelerated rate of gene family evolution compared to a random birth-and-death process. Comparing echolocating and non-echolocating species, we found clear contraction of OR genes in the former and significant support for expansion in the latter (Fig. 2, Supplementary Tables S15–S18). OR genes typically form the largest gene family in mammalian genomes and show high variation among species in terms of number, as well as in their degree of pseudogenization[32,33]. In New World fruit bats (family Phyllostomidae), variation in OR number appears to correlate with niche specialization[14], while contraction in the OR gene repertoire in the *M. brandtii* genome was suggested to reflect an evolutionary shift from olfaction to echolocation[12]. Our findings of significant contraction in the root of all bats as well as in echolocating forms, lend some support to this idea, perhaps pointing to a "trade-off" between sensory modalities, although wider sampling, including echolocating forms from the Yinpterochiroptera, is needed to confirm this trend (see also refs 16, 34) (Supplementary Tables S15 and S16).

Surveys of HOGs across 20 mammalian genomes also revealed expansion in up to 80 orthologous genes with links to immunity in bats (although this high turnover might arise from the difficulties of annotating highly divergent loci). Bats are well-known as reservoirs for a range of zoonotic diseases[35] and, as such, there have been a number of recent studies aimed at understanding the evolutionary dynamics of bat immunity genes[9,12,36,37]. Previous work has reported molecular adaptation in DNA repair loci and innate immune pathways in bat lineages[11,15]. Some immunity genes appear to have undergone expansion in bat genomes, including the leukocyte receptor complex (LRC) superfamily in *M. davidii*[11] and the *FBXO31* gene involved in ubiquitin-mediated degradation in *M. brandtii*[12]. In our study, specific HOGs showing high levels of gene gain in bats included several transmembrane receptors implicated in immune responses, such as *CD*-, *CEA*- and *CR*-like proteins, glycoproteins (e.g. *AZGP1*) and proteoglycans (e.g. *PRG*-like genes). Interestingly, previous studies of bats have also revealed several cases of contraction in immunity genes; these include killer-cell immunoglobulin like receptors (KIRs), killer cell lectin-like receptors (KLRs or Ly49 receptors)[11], IFN-α genes[37], and PYHIN genes[11,38]. Our analyses of HOGs revealed that the contraction of PYHIN, previously reported for *P. alecto* and *M. davidii*, is common to all eight bat species studied.

Across all five focal lineages tested in our study, the only significant GO enrichment for loci showing gene gain or loss was seen in the ancestral branch of Old World fruit bats, and involved the loss of immunity genes implicated in B cell activation and antigen binding (Fig. 2, Supplementary Table S17). However, looking at a wider taxonomic scale, it is noteworthy that the high inferred plasticity in numbers of immunity genes in bats was also seen in other mammalian lineages (Supplementary Tables S6, S7 and S8–S13, respectively). Some of these changes appeared to be associated with an accelerated rate of evolution, as revealed for immunoglobulin proteins based on our ML analyses of birth-and-death. Additional evidence is needed to establish whether differences in the gene complement are associated with variation in immune response to pathogens.

Conclusions

We show that bat genomes are highly plastic with respect to the turnover of protein-coding genes, but that the rate of gene turnover appears to be similar to that of their close relatives within the Laurasiatheria. A strong trend of gene loss in ORs in bats, and echolocating lineages in particular, was the opposite to the trend seen in both the non-echolocating bats and the carnivores examined, suggestive of a potential trade-off between olfaction and other senses in auditory specialists. These findings appeared to be robust to a range of proposed tree topologies for the relationships among laurasiatherian lineages.

Overall, our findings indicate that gene turnover tends to involve families with similar functional profiles, notably loci involved in immunity, regulation, metabolism and responses to stimulus. One possible explanation for this could be that the current status quo of GO annotation is insufficient or over-generic to allow rigorous tests of associations between gene family changes and biological functions. It could also be that only a small number of genes are essential for phenotypic adaptation, and that the signature of changes in these within the genome is masked by changes in larger highly dynamic families, such as OR and immunity related genes. Alternatively, if the evolution of novel phenotypes is indeed mediated via gene duplication and/or loss, then the underlying mechanism may not necessary pertain to specific gene targets but may operate through the same highly plastic groups of genes.

Methods

Inference of homologues across bat and mammal genomes. To investigate the evolutionary dynamics of gene gain and loss in bats, we first identifed homologous genes among bats and other closely related laurasiatherian mammals using the OMA standalone software package[17]. The OMA pipeline clusters homologous sequences from complete genomes in order to identify orthologous pairs of sequences and infer hierarchical orthologous groups ("HOGs") of genes that have descended for a common ancestral gene in a specified taxonomic range. OMA has been shown to be among the most reliable orthology inference methods—notably outperforming many tree-based methods (e.g. refs 39–41).

We augmented the two bat species contained in the May 2016 OMA database (*M. lucifugus* and *P. vampyrus*) with six bat proteomes retrieved from the GenBank (*E. fuscus*, *M. natalensis*, *M. brandtii*, *M. davidii*, *P. alecto* and *R. aegyptiacus*). As outgroups, we also included 12 other laurasiatherian mammals: *Ailuropoda melanoleuca* (giant panda), *Bos taurus* (cow), *Canis familiaris* (dog), *Equus caballus* (dog), *Erinaceus europaeus* (European hedgehog), *Felis catus* (cat), *Mustela furo* (ferret), *Ovis aries* (sheep), *Sorex araneus* (common shrew), *Sus scrofa* (pig), *Tursiops truncatus* (bottlenose dolphin) and *Vicugna pacos* (alpaca). Although additional genome and transcriptome data were available for bats at the time of the analysis (e.g. refs 13, 42–44), we decided to focus only on species whose genome was either Sanger sequenced or sequenced using high-throughput technologies at very high depth of coverage (>75X). Our reasoning was based upon previous findings showing that gene prediction and/or annotation errors could inflate estimates of gene turnover for taxa with low coverage draft genomes[30], a type of bias that we wanted to alleviate -or at least minimize- in downstream analyses. To this end, we also assessed gene annotation completeness of our 20 proteomic datasets using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software, based on a 3,300 gene set conserved across vertebrates[18].

If alternative protein isoforms were present for a given locus in any of the above proteomes, we only kept the first splicing variant (usually the longest). We predefined a species tree topology based on the most recent phylogenomic study of bats by Tsagkogeorga *et al.*[13], and ran the OMA pipeline under default parameters. To control that the inference of gene gains and losses was not biased by the predefined input phylogeny, we repeated our analysis using a tree topology estimated *de novo* directly from the proteomic data using OMA (parameter "SpeciesTree" set to "estimate" in the parameter file), as well as six proposed alternative species tree topologies for the diversification of laurasiatherian orders (Supplementary Figure S1). To map events of gene contraction and expansion at specific branches of the tree, the OMA hierarchical group output was parsed with HAM, a Python program developed by the authors available at http://lab.dessimoz.org/ham, inferring the placement of gene gains, duplications and losses using a parsimony criterion (with all types of events equally weighted).

Maximum likelihood estimation of the rate of gene turnover. To estimate the rate of HOG expansion and contraction in bats, we analysed the HOG data inferred from the OMA pipeline using the software CAFE 3[31]. For each HOG, we counted the number of genes present for each sampled genome in a given group of homologues, and converted these HOG counts at the tips of our tree into a CAFE formatted dataset. Again we used the species tree topology of Tsagkogeorga *et al.*[13] (consistent with the *de novo* tree inferred by OMA), together with divergence times taken from Meredith *et al.*[45] and from www.timetree.org to build the following reference species tree: (((*R. aegyptiacus*: 24, (*P. alecto*: 12, *P. vampyrus*: 12): 12): 42, (*M. natalensis*: 45, (*E. fuscus*: 25, (*M. davidii*: 14, (*M. lucifugus*: 9, *M. brandtii*: 9): 5): 11): 20): 21): 14, ((*F. catus*: 55, (*C. l. familiaris*: 46, (*M. p. furo*: 40, *A. melanoleuca*: 40): 6): 9): 24, (*E. caballus*: 78, (*V. pacos*: 65, (*S. scrofa*: 64, (*T. truncatus*: 56, (*O. aries*: 26, *B. taurus*: 26): 30): 8): 1): 13): 1): 1).

To accommodate errors potentially present in our HOGs dataset, we used the caferror.py script of the CAFE 3 software package, which estimates the error in a dataset without *a priori* information on the error distribution. We ran CAFE with the error estimation on HOGs with at least one gene present in the most recent common ancestor of Scrotifera (the clade uniting bats, carnivores and ungulates). We also used the option –s in caferror.py in order to obtain estimates of error for each of our 18 mammalian genomes separately (excluding insectivores used as outgroups) in addition to the average global error estimate across all species of our phylogeny.

Testing for accelerated evolution in bats. To assess the rate of gene turnover in bats as well as in other closely related mammals, we fitted to our data three birth-and-death models of gene family size evolution[46]: (i) a null model in which we defined a single global evolutionary rate of gene gain and loss (λ) across our tree; (ii) a two-lambda model, in which we specified a different rate of gene family turnover for the bat clade (λ_{bats}) compared to the rest of the mammalian phylogeny (λ_0); and (iii) a three-lambda model, in which we specified a different rate of gene family turnover for non-echolocating Old World fruit bats (suborder Yinpterochiroptera, λ_{owf}) and echolocating forms (suborder Yangochiroptera, λ_{yan}), and a third rate for the rest of the tree (λ_0). We ran all models five times to confirm convergence to a single global maximum, allowing for separate parameterization of the rate of gene birth (λ) and gene death (μ, where $\lambda \neq \mu$) or assuming one single parameter for the average turnover of genes within a HOG ($\lambda = \mu$). Analyses were repeated after correction for global error in the data, and after accounting for errors in each species separately. The p-value threshold in all runs was again specified at 0.01, and analyses were restricted to families with at least one gene in the root of the Scrotifera clade in the reference tree. To account for multiple testing, we followed the Benjamini & Hochberg's procedure[47] that corrects the false discovery rate (FDR). To compare estimated rates of gene change in bats to other mammalian lineages, analyses were repeated with the Carnivora and Cetartiodactyla as focal groups, respectively (data not shown).

To assess the significance of the observed HOG size differences among bats and other laurasiatherian genomes, we generated 5,000 simulated datasets under the global λ estimate after correcting for each individual species' error (using the command *genfamily* in CAFE). This dataset was subsequently used to build a null distribution of likelihood ratios under the model of a global lambda versus one with two-lambda values assuming

two independent rates, one for bats and one for the other mammals. Differences in the model fit were considered significant if they fell outside of the 95% of the null distribution.

Functional annotation. To annotate HOGs in terms of proteins and their putative role in species biology, we first interrogated protein information available in UniProtKB database[26]. We selected one representative protein for each HOG, preferably from the cow when present, else from dog, cat or horse, as these species appeared to have better annotation records compared to the rest of our sampled mammals. Then, for a given list of HOGs of interest, we used the collected protein identifiers as queries in UniProtKB to infer functional groups based on GO terms, as well as based on keywords from protein annotations.

Second, we tested for enrichment in GO terms of HOGs showing gene gains and losses in the respective MRCAs of all bats, Old World fruit bats, echolocating bats, as well as in the MRCAs of Ferungulata, carnivores, and even-toed ungulates and cetaceans (Cetartiodactyla). Again, we sampled one cow sequence from each HOG, such as each HOG was linked to a cow gene and its associated GO terms as a proxy for the functional role of its members. In addition, based on the same set of identifiers we used Ensembl Biomart[48] to retrieve human gene homologues, which are better annotated than other mammalian genes and thus could potentially offer better insights into the functional profile of our gene sets.

The most current associations of cow and human genes with GO terms were download from the Gene Ontology website (October 2016), and GO enrichment analyses were performed based on Fisher's exact test, using the Python package GOATOOLS (https://github.com/tanghaibao/Goatools). The background population of HOGs for the GO analyses was defined separately in each test, and included all HOGs present in the parental node of the branch tested for enrichment. Finally, resulted p-values were adjusted for multiple testing using Bonferroni, Sidak, and Holm corrections, also implemented in GOATOOLS[49].

Correlation analysis between λ and genome size. The genome size (C-value) for each of our sampled species was obtained from the Animal Genome Size Database[7]. If the exact species did not have an entry in the database, we used an expected C-value calculated from the average C-value from other species of the same genus. Nonparametric Spearman correlation tests between genome size values, and estimates of average gene expansion, gene gain and gene loss were performed in R. To account for the phylogenetic history of the compared species, we used the comparative method phylogenetic independent contrasts (PIC) implemented in the PHYLIP package[50].

References

1. Ohno, S. *Evolution by gene duplication* (Springer-Verlag, 1970).
2. Zhang, J. Z. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292–298, doi:10.1016/S0169-5347(03)00033-8 (2003).
3. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *P Roy Soc B-Biol Sci* **279**, 5048–5057, doi:10.1098/rspb.2012.1108 (2012).
4. Zhang, F., Gu, W. L., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genom Hum G* **10**, 451–481, doi:10.1146/annurev.genom.9.081307.164217 (2009).
5. Ames, R. M., Money, D., Ghatge, V. P., Whelan, S. & Lovell, S. C. Determining the evolutionary history of gene families. *Bioinformatics* **28**, 48–55, doi:10.1093/bioinformatics/btr592 (2012).
6. Rho, M. *et al.* Independent mammalian genome contractions following the KT boundary. *Genome Biol Evol* **1**, 2–12, doi:10.1093/gbe/evp007 (2009).
7. Gregory, T. R. Animal genome size database (2016).
8. Smith, J. D. L., Bickham, J. W. & Gregory, T. R. Patterns of genome size diversity in bats (order Chiroptera). *Genome* **56**, 457–472, doi:10.1139/gen-2013-0046 (2013).
9. Zhang, G. J. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320, doi:10.1126/science.1251385 (2014).
10. Zhang, Q. & Edwards, S. V. The evolution of intron size in amniotes: a role for powered flight? *Genome Biol Evol* **4**, 1033–1043, doi:10.1093/gbe/evs070 (2012).
11. Zhang, G. J. *et al.* Comparative analysis of bat genomes provides Insight into the evolution of flight and immunity. *Science* **339**, 456–460, doi:10.1126/science.1230835 (2013).
12. Seim, I. *et al.* Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nat Commun* **4**, doi:ARTN221210.1038/ncomms3212 (2013).
13. Tsagkogeorga, G., Parker, J., Stupka, E., Cotton, J. A. & Rossiter, S. J. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr Biol* **23**, 2262–2267, doi:10.1016/j.cub.2013.09.014 (2013).
14. Hayden, S. *et al.* A cluster of olfactory receptor genes linked to frugivory in bats. *Mol Biol Evol* **31**, 917–927, doi:10.1093/molbev/msu043 (2014).
15. Xu, H. H. *et al.* Multiple bursts of pancreatic ribonuclease gene duplication in insect-eating bats. *Gene* **526**, 112–117, doi:10.1016/j.gene.2013.04.035 (2013).
16. Zhao, H. B., Xu, D., Zhang, S. Y. & Zhang, J. Z. Genomic and genetic evidence for the loss of umami taste in bats. *Genome Biol Evol* **4**, 73–79, doi:10.1093/gbe/evr126 (2012).
17. Altenhoff, A. M. *et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* **43**, D240–D249, doi:10.1093/nar/gku1158 (2015).
18. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, doi:10.1093/bioinformatics/btv351 (2015).
19. Waddell, P. J., Okada, N. & Hasegawa, M. Towards resolving the interordinal relationships of placental mammals. *Syst Biol* **48**, 1–5 (1999).
20. Prasad, A. B., Allard, M. W., Program, N. C. S. & Green, E. D. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol* **25**, 1795–1808, doi:10.1093/molbev/msn104 (2008).
21. O'Leary, M. A. *et al.* The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339**, 662–667, doi:10.1126/science.1229237 (2013).
22. Nishihara, H., Hasegawa, M. & Okada, N. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc Natl Acad Sci USA* **103**, 9929–9934, doi:10.1073/pnas.0603797103 (2006).
23. McCormack, J. E. *et al.* Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res* **22**, 746–754, doi:10.1101/gr.125864.111 (2012).
24. Madsen, O. *et al.* Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610–614, doi:10.1038/35054544 (2001).

25. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482, doi:10.1038/nature10530 (2011).
26. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database-Oxford*, doi:ARTNbar00910.1093/database/bar009 (2011).
27. Demuth, J. P. & Hahn, M. W. The life and death of gene families. *Bioessays* **31**, 29–39, doi:10.1002/bies.080085 (2009).
28. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**, 97–108, doi:10.1038/nrg2689 (2010).
29. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**, 121–152, doi:10.1146/annurev.genet.39.073003.112240 (2005).
30. Denton, J. F. *et al.* Extensive error in the number of genes inferred from draft genome assemblies. *Plos Comput Biol* **10**, doi:ARTNe100399810.1371/journal.pcbi.1003998 (2014).
31. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30**, 1987–1997, doi:10.1093/molbev/mst100 (2013).
32. Hayden, S. *et al.* Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res* **20**, 1–9, doi:10.1101/gr.099416.109 (2010).
33. Niimura, Y. & Nei, M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One* **2**, e708, doi:10.1371/journal.pone.0000708 (2007).
34. Zhao, H. *et al.* The evolution of color vision in nocturnal mammals. *Proc Natl Acad Sci USA* **106**, 8980–8985, doi:10.1073/pnas.0813201106 (2009).
35. Wynne, J. W. & Wang, L. F. Bats and viruses: friend or foe? *PLoS Pathog* **9**, e1003651, doi:10.1371/journal.ppat.1003651 (2013).
36. Daugherty, M. D., Young, J. M., Kerns, J. A. & Malik, H. S. Rapid evolution of PARP genes suggests a broad role for ADP-ribosylation in host-virus conflicts. *PLoS Genet* **10**, e1004403, doi:10.1371/journal.pgen.1004403 (2014).
37. Zhou, P. *et al.* Contraction of the type I IFN locus and unusual constitutive expression of IFN-alpha in bats. *Proc Natl Acad Sci USA* **113**, 2696–2701, doi:10.1073/pnas.1518240113 (2016).
38. Cridland, J. A. *et al.* The mammalian PYHIN gene family: phylogeny, evolution and expression. *BMC Evol Biol* **12**, 140, doi:10.1186/1471-2148-12-140 (2012).
39. Afrasiabi, C., Samad, B., Dineen, D., Meacham, C. & Sjolander, K. The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res* **41**, W242–248, doi:10.1093/nar/gkt399 (2013).
40. Boeckmann, B., Robinson-Rechavi, M., Xenarios, I. & Dessimoz, C. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform* **12**, 423–435, doi:10.1093/bib/bbr034 (2011).
41. Altenhoff, A. M. *et al.* Standardized benchmarking in the quest for orthologs. *Nat Methods* **13**, 425–430, doi:10.1038/nmeth.3830 (2016).
42. Francischetti, I. M. *et al.* The "Vampirome": Transcriptome and proteome analysis of the principal and accessory submaxillary glands of the vampire bat *Desmodus rotundus*, a vector of human rabies. *J Proteomics* **82**, 288–319, doi:10.1016/j.jprot.2013.01.009 (2013).
43. Lee, A. K. *et al. De novo* transcriptome reconstruction and annotation of the Egyptian rousette bat. *BMC Genomics* **16**, 1033, doi:10.1186/s12864-015-2124-x (2015).
44. Shaw, T. I. *et al.* Transcriptome sequencing and annotation for the Jamaican fruit bat (*Artibeus jamaicensis*). *PLoS One* **7**, e48472, doi:10.1371/journal.pone.0048472 (2012).
45. Meredith, R. W. *et al.* Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524, doi:10.1126/science.1211028 (2011).
46. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271, doi:10.1093/bioinformatics/btl097 (2006).
47. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
48. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res* **42**, D749–755, doi:10.1093/nar/gkt1196 (2014).
49. Tang, H. X. *et al.* GOATOOLS: Tools for gene ontology (2015).
50. PHYLIP (Phylogeny Inference Package) version 3.5c (Department of Genetics, University of Washington, Seattle, 1993).

Acknowledgements

We thank Kalina Davies, Michael McGowen, Joshua Potter and Kimberley Warren for helpful advice and discussions. Analyses were performed with the assistance of SBCS-Informatics (http://informatics.sbcs.qmul.ac.uk) and the EPSRC-funded MidPlus cluster at Queen Mary University of London. SJR was funded by the European Research Council (ERC 1076 Starting grant 310482), and CD by the Swiss National Science Foundation (grant 150654).

Authors' Contributions

S.J.R. and G.T. conceived the project and designed the study. G.T. analysed the data, with input from C.D., S.M. and S.J.R. G.T. and S.J.R. drafted the manuscript. All other authors assisted in revising the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00132-9

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

OMA standalone: orthology inference among public and custom genomes and transcriptomes

Adrian M Altenhoff[1,2], Jeremy Levy[3,4], Magdalena Zarowiecki[5], Bartłomiej Tomiczek[4], Alex Warwick Vesztrocy[1,4], Daniel A Dalquen[4], Steven Müller[4], Maximilian J, Telford[4], Natasha M Glover[1,6,7], David Dylus[1,6,7], Christophe Dessimoz[1,4,6,7,8,*]

[1]Swiss Institute of Bioinformatics, Genopode Building, Lausanne, Switzerland
[2]Department of Computer Science, ETH Zurich, Lausanne, Switzerland
[3]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK
[4]Centre for Life's Origins and Evolution, Department of Genetics, Evolution & Environment, University College London, London, UK
[5]Genomics England, Queen Mary University of London, Dawson Hall, London, EC1M 6BQ
[6]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[7]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
[8]Department of Computer Science, University College London, London, UK

*Corresponding author: Christophe.Dessimoz@unil.ch

Abstract

Genomes and transcriptomes are now typically sequenced by individual labs, but analysing them often remains challenging. One essential step in many analyses lies in identifying orthologs—corresponding genes across multiple species—but this is far from trivial. The OMA (Orthologous MAtrix) database is a leading resource for identifying orthologs among publicly available, complete genomes. Here, we describe the OMA pipeline available as a standalone program for Linux and Mac. When run on a cluster, it has native support for the LSF, SGE, PBS Pro, and Slurm job schedulers and can scale up to thousands of parallel processes. Another key feature of OMA standalone is that users can combine their own data with existing public data by exporting genomes and pre-computed alignments from the OMA database, which currently contains over 2100 complete genomes. We compare OMA standalone to other methods in the context of phylogenetic tree inference, by inferring a phylogeny of the Lophotrochozoa, a challenging clade within the Protostomes. We also discuss other potential applications of OMA standalone, including identifying gene families having undergone duplications/losses in specific clades, and identifying potential drug targets in non-model organisms. OMA Standalone is available at http://omabrowser.org/standalone under the permissible open source Mozilla Public License Version 2.0.

Introduction

The sequencing revolution is yielding a flood of genomes and transcriptomes, with thousands already sequenced and many more underway (Pagani et al., 2012). A powerful way of characterising newly sequenced genes is to compare them with evolutionarily related genes—in particular with orthologs in other species (Dessimoz et al., 2012; Forslund et al., 2017; Sonnhammer et al., 2014). In this way, experimental knowledge from model organisms can be propagated to non-model organisms. Elucidation of orthology and paralogy relationships is also essential to reconstruct species trees, to better understand the mechanics of gene/genome evolution, to study adaptation, or to pinpoint the emergence of new gene functions (Gabaldón and Koonin, 2013).

The importance of determining orthology has led to the development of many inference methods and associated databases (reviewed in Altenhoff and Dessimoz, 2012). Some of the best established orthology resources include EggNOG (Huerta-Cepas et al., 2016), Ensembl Compara (Zerbino et al., 2018), Inparanoid (Sonnhammer and Östlund, 2015), MBGD (Uchiyama et al., 2012), OrthoDB (Zdobnov et al., 2017), OrthoMCL (Chen et al., 2006), Panther (Mi et al., 2017), PhylomeDB (Huerta-Cepas et al., 2014), and OMA (Altenhoff et al., 2017).

Key distinctive features of OMA are the high specificity of its inference pipeline (Afrasiabi et al., 2013; Altenhoff and Dessimoz, 2009; Boeckmann et al., 2011; Linard et al., 2011), the feature-rich web and programmatic interfaces, large size and taxonomic breadth of its precomputed data (currently 2167 genomes), its regular update schedule of 2 releases per year, and its sustained development over the last 13 years. The algorithms underlying the OMA pipeline have been described and validated in multiple publications (Altenhoff et al., 2013; Dessimoz et al., 2006, 2005; Roth et al., 2008; Train et al., 2017). The quality of OMA is corroborated by a recent community experiment, which highlighted the high specificity of orthologs predicted by the OMA pipeline (Altenhoff et al., 2016).

With genome and transcriptome sequencing rapidly becoming a commodity, there is an increasing need to analyse custom user data. Here, we present OMA standalone, an open-access software implementation of the OMA pipeline for Linux and Mac. We first outline some of the key features of OMA standalone. In the second part, we demonstrate the usefulness of OMA standalone in the context of species tree inference, by comparing its performance with state-of-the-art alternatives on the challenging Lophotrochozoa phylogeny.

Results

We first highlight the defining features of OMA standalone, then turn to the phylogeny of the Lophotrochozoa, which we infer from orthologs inferred by OMA in comparison with alternative methods.

OMA standalone software

OMA standalone takes as input the coding sequences of genomes or transcriptomes, in fasta format. The recommended input type is amino-acid sequences, but OMA also supports nucleotide sequences. With amino-acid sequences, users can combine their own data with publicly available genomes from the OMA database, including precomputed all-against-all comparisons, using the export function on the OMA website (http://omabrowser.org/export).

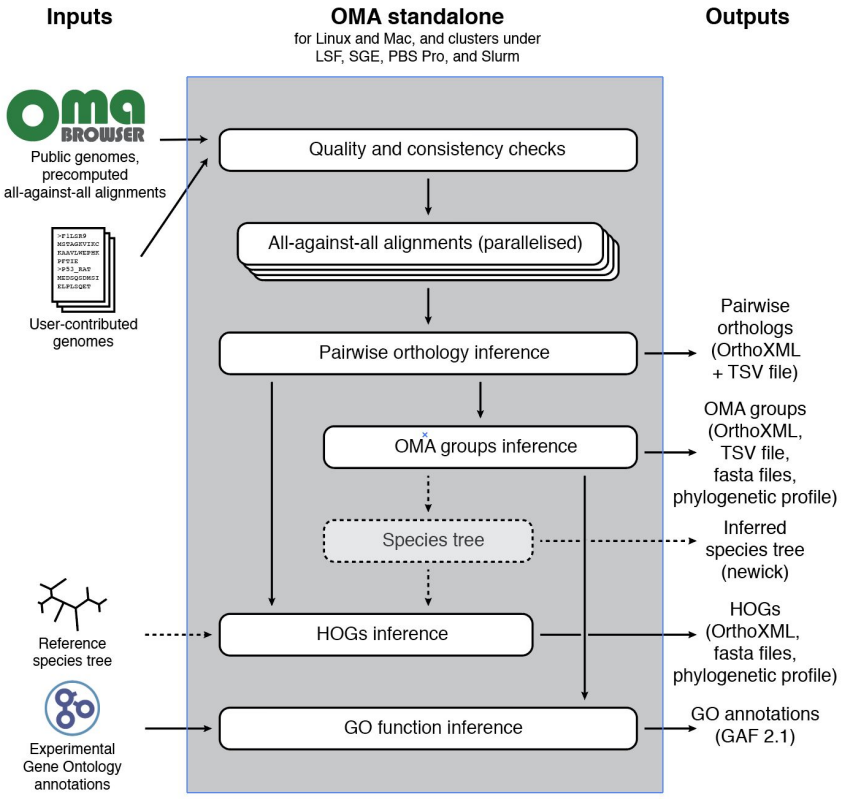


Fig. 1. Conceptual overview of the OMA standalone software. Dotted arrows indicate alternative steps (reference species tree either specified as input or inferred from the data).

OMA standalone produces several types of output (also summarised in Fig. 1):

1. **Pairwise orthologs** and their subtypes (1:1, 1:many, many:1, many:many orthology). These orthologs are useful when comparing pairs of species at a time, or to identify orthologs to specific genes of interest
2. **OMA groups.** These are sets of genes for which all pairs are inferred to be orthologous. These groups are inferred as cliques (fully connected subgraphs) of pairwise orthologs. These groups are not necessarily one-to-one orthologs, but being inferred without assuming a species tree, they are particularly useful to identify marker genes for phylogenetic reconstruction.

3. **Hierarchical orthologous groups (HOGs).** These groups are defined for every internal node of the (rooted) species tree; each HOG contains the genes that are inferred to have descended from a common ancestral gene among the species attached to that internal node. Consider for instance gene ADH1, which duplicated within the primates (Carrigan et al., 2012): At the level of the last primate common ancestor, all genes that have descended from the ancestral ADH1 belong to the same HOG. However, at the level of the common ancestor of all the great apes, because ADH1 had at this point already duplicated into ADH1a, ADH1b, and ADH1c, these ancestral genes define 3 HOGs. The HOGs are stored in the standard OrthoXML format (Schmitt et al., 2011).
4. **Gene Ontology annotations.** OMA standalone annotates the input sequences with Gene Ontology annotations by propagating high-quality annotations across orthologs (Altenhoff et al., 2015). The annotations are provided in the standard GO Annotation File Format 2.1 (http://geneontology.org/page/go-annotation-file-format-20).
5. **Phylogenetic profiling.** Orthology is also used to build phylogenetic profiling—patterns of presence and absence of genes across species (Pellegrini et al., 1999). We provide two forms of output: a binary matrix with species as rows and OMA groups as columns, indicating patterns of presence or absence of genes in each group; a count matrix with species as columns and HOGs as rows, indicating the number of genes in each HOG.

OMA standalone supports parallel computation of the all-against-all sequence comparison phase. This phase, which computes Smith-Waterman (1981) alignments followed by pairwise maximum likelihood distance estimation for all significant pairs (Roth et al., 2008), is by far the most time-consuming step of the algorithm. To fully exploit parallelism, alignments are performed using single instruction multiple data (SIMD) instructions (Szalkowski et al., 2008) on multiple cores. OMA standalone natively supports common cluster schedulers—LSF, SGE, PBS, and Slurm—and has been successfully run with several thousand jobs in parallel. Figure 2 shows typical runtimes and memory usage for datasets of various sizes.

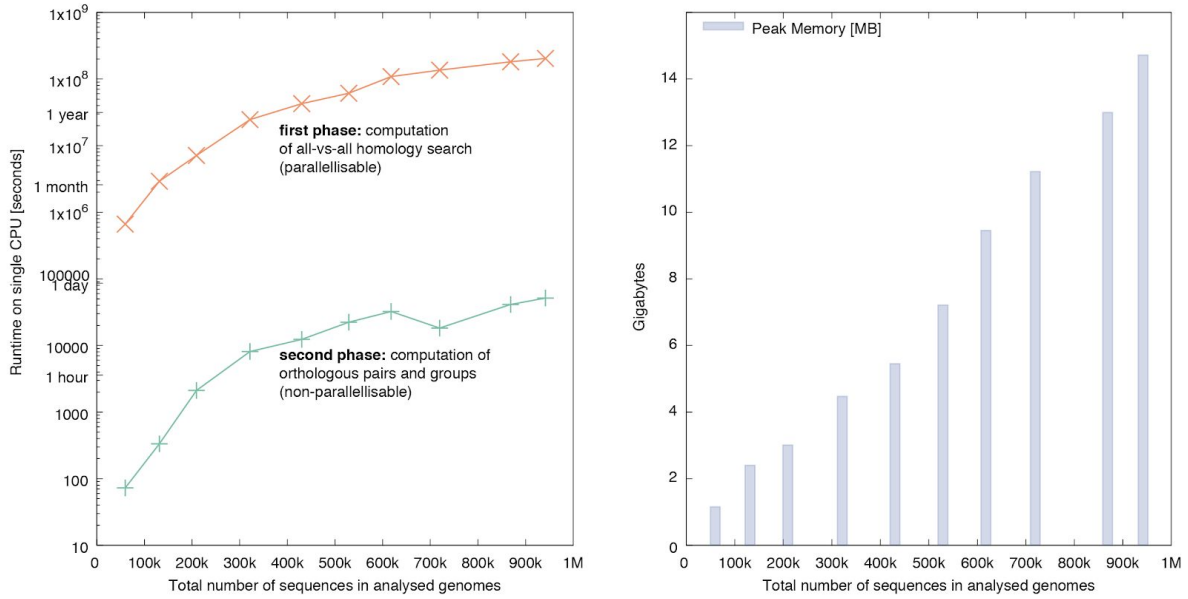


Fig 2: Resource measurements for various datasets of increasing sizes as total number of protein sequences. The datasets have been sampled from the public OMA Browser to maintain a constant composition of 20% fungi, 10% archaea, 10% plants, 20% metazoa and 40% bacteria genomes. **Left**: Runtime of the all-against-all phase (orange) on a single CPU, and the inference of the orthologous pairs and various groups (green). **Right**: Peak memory usage of OMA standalone in gigabytes.

Application: the phylogenetic relationships within the Lophotrochozoa

Resolving the relationships of ancient lineages is a major challenge for molecular phylogenetics. Although some aspects of the phylogeny of the major animal clades are well resolved, the relative positions of the deeper lying clades are often disputed. The construction of large phylogenomic supermatrices, has been the method of choice for resolving the deepest nodes in the tree of life (Dunn et al., 2008; Egger et al., 2015; Fernández et al., 2014; Hejnol et al., 2009).

Fundamental to the analyses of phylogenetic relationships is the use of sequences which have descended from a single common gene in their last common ancestor, that is, orthologous sequences. Ensuring that we correctly infer orthologs is therefore vital if we are to reconstruct difficult to resolve phylogenies. The limitations of automated orthology and paralogy prediction methods with regards to phylogenetic analysis have previously been highlighted (Philippe et al., 2011b); simplistic orthology inference methods may miss orthologs (Dalquen and Dessimoz, 2013) or erroneously identify as orthologs, paralogous pairs of genes that result from differential gene losses (Dessimoz et al., 2006).

One notoriously difficult to resolve phylogeny is that of the Lophotrochozoa (Kocot, 2016), a clade of animals positioned sister to the Ecdysozoa, within the protostomes. The Lophotrochozoa contains about ten different phyla, each of which is clearly monophyletic, but the relationships between these

phyla are far from clear, with many different topologies having been supported by different analyses.The inference is that the phyla are likely to have emerged in an ancient and rapid radiation resulting in weak phylogenetic signal for interphylum relationships. These circumstances make the solving of this problem particularly difficult and mean the use of accurately identified orthologs is particularly significant.

We used OMA standalone to identify orthologous marker genes among the proteomes of 19 lophotrochozoans and, as outgroups, 4 deuterostomes, 4 ecdysozoans, and 3 non-bilaterians (see Material and Methods). As a basis of comparison, we also repeated the analysis using orthology inference pipelines based on OrthoMCL (Li et al., 2003), BUSCO (Simão et al., 2015), and HaMStR (Ebersberger et al., 2009). Like OMA, these methods do not require prior specification of a species tree, are available as standalone programs and have all been used in phylogenetic analyses previously. Species trees were then constructed using these orthologs with both maximum likelihood and Bayesian tree reconstruction packages, RAxML (Stamatakis, 2014) and PhyloBayes (Lartillot et al., 2013), on the resultant supermatrices.

We first consider the amount of orthology information recovered by the various methods. OMA inferred 2,162 orthologous groups containing 15 or more species (Figure 3a). By comparison, HaMStR pipelines inferred 1,192 orthologous groups, the OrthoMCL pipeline inferred 484 orthologous groups, and BUSCO inferred 384 orthologous groups. Although OMA overall identifies more orthologous genes than other methods, it infers fewer larger groups than HaMStR and OrthoMCL. The OMA algorithm is known for having higher precision but lower recall than most other methods (Altenhoff et al., 2016). Still, in terms of total number of characters in supermatrices, OMA standalone yields a larger matrix (i.e. alignment columns) than the other methods (Figure 3b).

Using the aligned sets of orthologs identified in the previous step, we reconstructed species trees using Maximum Likelihood (RaXML, LG+I model) and Bayesian analysis (PhyloBayes, CAT+GTR+G4) on supermatrices which had been filtered to include only alignment columns with at least 60% site occupancy.

With OMA, both the RAxML tree and the Phylobayes tree had high branch support values. The RAxML tree had bootstrap support of 100 for each branch, except for five. The Deuterostomes were recovered with bootstrap support of 89, whilst the Lophotrochozoa, with the exception of Rotifera, were recovered with bootstrap support of 92. Similarly, the PhyloBayes tree had branch posterior probabilities of 1 across the tree apart from the Lophotrochozoa clade, with a posterior probability of 0.82.

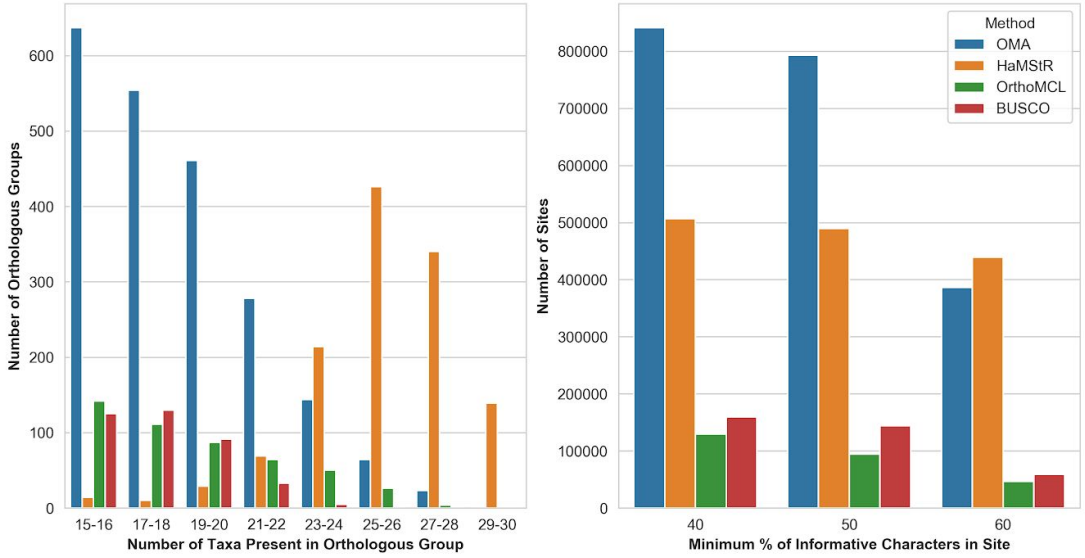


Fig. 3: Comparison of amount of orthologous data inferred by the different pipelines. A: OMA infers more orthologous groups than other methods; the groups inferred by HaMStR are considerably larger on average than for the other methods. B: The resulting supermatrix (concatenated alignment over all orthologous groups) has most sites for OMA whether the minimum site occupancy threshold is 40% or 50%.

The tree inferred using the ML inference method found that the Rotifera (Adineta ricciae, Brachionus plicatilis) are grouped with the Nematoda (Caenorhabditis elegans, Pristionchus pacificus), as part of the ecdysozoans. This is in disagreement with the current consensus (Giribet and Edgecombe, 2017). By contrast, the tree constructed using Bayesian inference found the Rotifera to be sister to the rest of the Lophotrochozoa, in agreement to recent studies (Egger et al., 2015; Philippe et al., 2011a). The discrepancy in the ML tree is likely due to the long branched Rotifera being attracted to the long branched Nematoda—a problem to which PhyloBayes under the CAT model has been previously shown to be more robust (Lartillot et al., 2013).

Both the ML and Bayesian trees found the rest of the Lophotrochozoa to consist of two monophyletic groups. The first group comprises of the Gastrotricha (Mesodasys laticaudatus), and the Platyhelminthes (flatworms). This relationship is consistent with recent studies (Dunn et al., 2008; Edgecombe et al., 2011; Laumer et al., 2015; Struck et al., 2014). Because of their primitive nature, with characteristics such as having no body cavity, no respiratory organs, and having only a single opening for both the intake of nutrients and excretion of waste, they were originally thought to be amongst the more primitive Bilateria, until molecular studies on 18S rDNA sequence data was carried out, placing them within the protostomes (Baguñà and Riutort, 2004). Authors now divide the Platyhelminthes into the Catenulida, with currently no known synapomorphies, and the Rhabditophora, which has uniting characteristics such as the presence of lamellated rhabdites, a common structure of the epidermis (Egger et al., 2015; Laumer et al., 2015). Our ML and Bayesian

trees corroborated this, and found the Catenulida (Catenulida sp.) to be sister to Rhabditophora (Macrostomum lignano, Echinoplana celerrima, Microdalyellia schmidti, Monocelis, Schmidtea mediterranea).

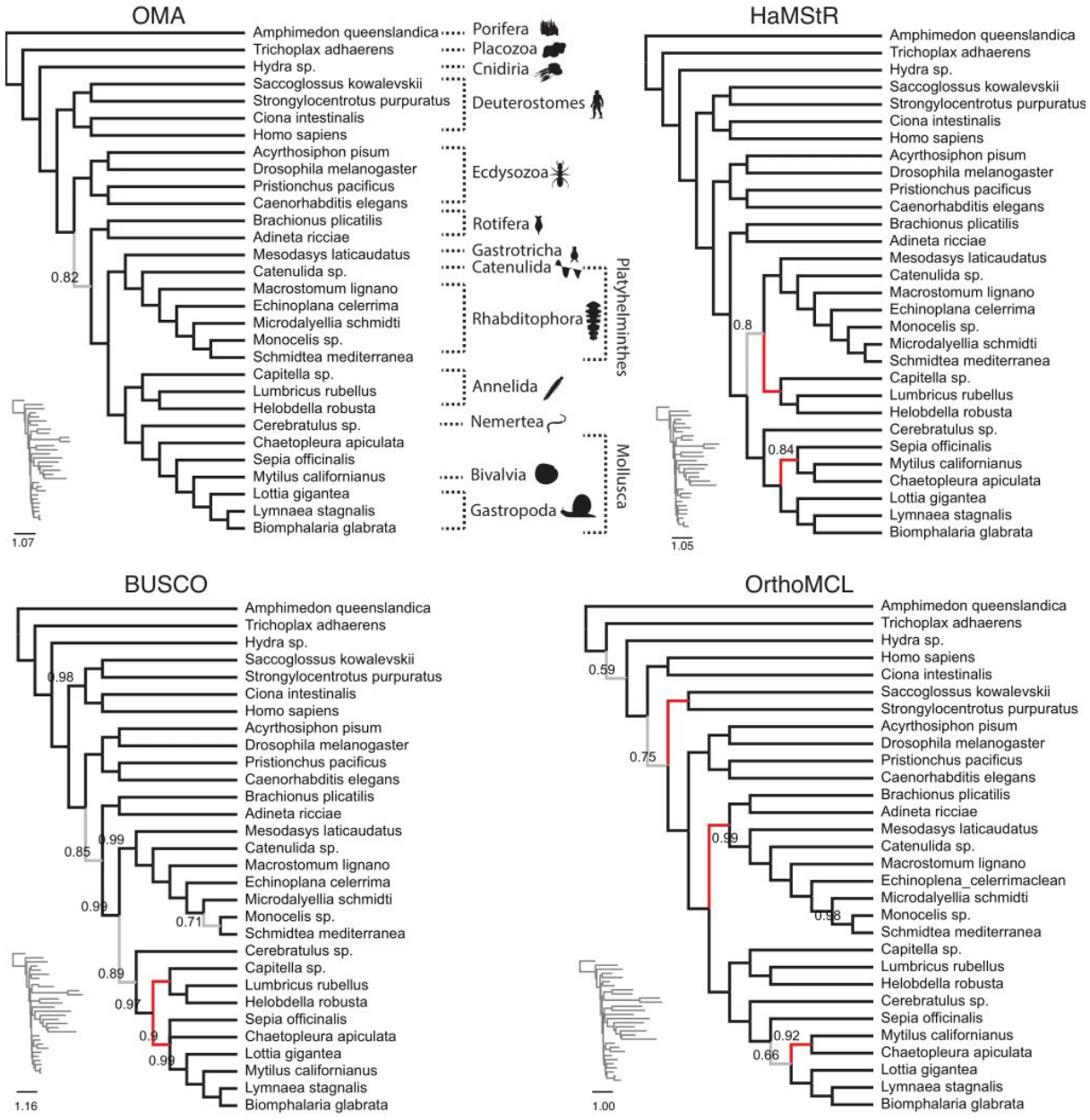


Fig. 4: Comparison of trees obtained using PhyloBayes with the CAT-GTR-G4 model for different datasets. OMA tree is in congruence with published results (see main text). Branches which are at odds with the literature are in red; else they are displayed in grey (posterior probability < 0.95) or else in black. Only posterior probabilities below 1 are displayed.

Table 2: Summary of support for major clades in trees obtained using the different methods. P indicates presence of clade in PhyloBayes trees (bold: posterior probability >=0.95). L indicates presence of clade in maximum likelihood tree (bold: branch support >=0.95)

Hypothesis	OMA	OrthoMCL	HamSTR	BUSCO
Monophyly of Lophotrochozoa ((Dunn et al., 2008; Kocot et al., 2017; Telford et al., 2015))	**P** -	**P** -	**P** -	**P** -
Gastropoda sister to Bivalvia ((Kocot et al., 2011))	**P** **L**	- L	- -	- -
Annelida to be sister to (Mollusca + Nemertea) (Egger et al., 2015)	**P** **L**	P **L**	- **L**	P **L**
Monophyly of Deuterostomes	**P** L	- L	**P** **L**	**P** **L**
Rotifera sister to rest of Lophotrochozoa (Laumer et al., 2015)	P -	- -	**P** -	P -
Catenulida sister to Rhabditophora (Egger et al., 2015)	**P** **L**	**P** L	**P** **L**	**P** **L**
Monophyly of Annelida	**P** **L**	**P** **L**	**P** **L**	**P** **L**
Sister Clade of Gastrotricha to Platyhelminthes (Egger et al., 2015; Laumer et al., 2015)	**P** L	**P** -	**P** **L**	**P** -
Total majority outcomes	**8** (PhyloBayes) **6** (RaxML)	5 (PhyloBayes) 5 (RaxML)	6 (PhyloBayes) 5 (RaxML)	7 (PhyloBayes) 4 (RaxML)

Within the Rhabditophora, the most basal branches are those of the Macrostomorpha (Macrostomum lignano), followed by the Polycladida (Echinoplana celerrima), also in agreement with recent studies (Egger et al., 2015; Laumer et al., 2015). We see a disagreement between the ML and Bayesian tree

topologies regarding the rest of the Rhabditophora. The ML tree inferred the Proseriata (Monocelis sp.) to be more basal than both the Rhabdocoela (Microdalyellia schmidti) and the Acentrosomata (Schmidtea mediterranea). This is in disagreement with recent analyses (Egger et al., 2015; Laumer et al., 2015), which places the Rhabdocoela as the most basal, followed by Proseriata and then Acentrosomata. The tree found through Bayesian inference agrees with the pedlishub phylogenies, however. As with the placement of the Rotifera within the Ecdysozoans under the ML analysis, this is possibly due to the ML inference method being more susceptible to Long Branch Attraction artefacts than the Bayesian, leading the long branched Rhabdocoela to be attracted to the long branched Acentrosomata in this framework.

The second monophyletic group found within the rest of the Lophotrochozoa contains the Annelida (Lumbricus rubellus Helobdella robusta Capitella sp.), worms, the Mollusca (Biomphalaria glabrata, Lymnaea stagnalis, Lottia gigantea, Mytilus californianus, Sepia officinalis, Chaetopleura apiculata), the largest marine phylum, and Nemertea (Cerebratulus sp.), also known as ribbon worms or proboscis worms, to form the Trochozoa (Dunn et al., 2014). However, there is disagreement on the positioning of these clades within the group (Dunn et al., 2008; Laumer et al., 2015; Struck et al., 2014; Struck and Fisse, 2008). Both tree reconstruction methods find the Gastropoda (Lottia gigantea, Lymnaea stagnalis, Biomphalaria glabrata) to be sister to the Bivalvia (Mytilus californianus). Both methods also found the Annelida to be sister to (Mollusca + Nemertea), with high support (posterior probability of 1 and bootstrap of 100).

By contrast, trees obtained from other orthology pipelines had more unresolved nodes and/or more discrepancies with the literature (Figure 4; table 1).

The BUSCO Bayesian tree had slightly less support throughout than the OMA tree, although only had one branch with support of less than pp=0.80. The relationship between the Poreriata, Rhabdocoela and the Acentrosomata agrees with the OMA Bayesian tree, as does the relationship between the Gastrotricha and the Platyhelminthes. However, the BUSCO tree indicates Gastropoda to be paraphyletic with high support (pp=0.99), with Lottia gigantea to be more basal to the Bivalvia and the rest of the Gastropoda. This is in contrast to both the OMA tree and other studies (Dunn et al., 2008; Struck et al., 2014). The BUSCO tree found the Nemertea as sister to (Annelida + Mollusca), with a support value of pp=0.89. This is in disagreement with current consensus, and the OMA tree (Dunn et al., 2008; Laumer et al., 2015; Struck et al., 2014).

The HaMStR tree had high support throughout, but differed from the OMA tree. The HaMStR method placed the Sepia officinalis, Mytilus californianus and the Chaetopleura apiculata in a clade together, sister to the Gastropoda. This is in disagreement with (Kocot et al., 2011) and the OMA trees, which place the Polyplacophora (Chaetopleura apiculata) to be the most basal, followed by the Cephalopoda (Sepia officinalis), with the Bivalvia sister to the Gastropoda. The Bayesian tree also fails to recover the Trochozoa, placing the Annelida with the (Platyhelminthes+Gastrotricha), as opposed to full support found in the OMA tree.

The OrthoMCL trees had the most issues, with the lowest support values. Deuterostomes, comprising of a well established relationship between the Chordates and the Ambulacraria (Philippe et al., 2011a), are paraphyletic in the Phylobayes tree, which places the Chordates (Ciona intestinalis, Homo sapiens) basal to the Ambulacraria (Strongylocentrotus purpuratus, Saccoglossus kowalevskii), with the latter sister to the Protostomes with pp=0.75. The Rotifera were incorrectly placed as sister to (Gastrotricha + Platyhelminthes) with full support. This is in disagreement with both the OMA tree and recent studies. The tree was able to correctly infer the (Mollusca + Nemertea) relationship with full support. Within the Mollusca, in contrast to the OMA tree, the Bayesian tree inferred the Sepia officinalis to be the most basal, with Chaetopleura apiculata and Mytilus californianus forming a clade sister to the rest of the Mollusca. However, this has low support with pp=0.66 for the Bayesian tree.

Discussion and outlook

OMA standalone enables researchers to infer high-quality orthologs among genomes or transcriptomes, on public and in house data. It runs on a wide range of hardwares, from a single computer to large clusters with thousands of parallel processes.

On the Lophotrochozoa dataset, compared with other approaches, OMA yielded more orthologous information for phylogenetic species tree inference and resulted in better resolved trees which are more consistent with the existing literature.

OMA standalone was also successful used to analyse centipedes (Fernández et al., 2014), arachnids (Fernández and Giribet, 2015; Sharma et al., 2014), assassin flies (Dikow et al., 2017), scorpions (Sharma et al., 2015), spiders (Garrison et al., 2016), flatworms (Egger et al., 2015; Laumer et al., 2015), tapeworms (Tsai et al., 2013), or Archaea (Williams et al., 2017).

Beyond species tree inference, OMA can also be used to pinpoint the emergence of gene families in evolution, an approach that is sometimes referred to as phylostratigraphy (Domazet-Lošo et al., 2007). Conventional approaches work by considering all the genes annotated in a species of reference, and performing BLAST searches against increasingly distant sets of taxa. The point at which no homolog can be found is inferred to immediately precede the emergence of the gene. However, such an approach does not differentiate between orthologs and paralogs, and thus has a limited resolution in terms of subfamilies. Alternatively, it is possible to extract more fine-grained information from reconciled gene trees (Huerta-Cepas et al., 2014; e.g. Vilella et al., 2008), but this is computationally demanding and there is a lack of tools to perform such analyses on custom data.

By inferring high-quality hierarchical orthologous groups, OMA standalone provides a way to map gene emergence, gene duplication, and gene loss onto species phylogenies. For instance, OMA standalone has been used to contrast gene families that have expanded and contracted in the common ancestors of echolocating and non-echolocating bats. The emergence of echolocation

coincides with a decrease in chemosensory genes, while secondary loss of echolocation coincides with an increase in chemosensory genes (Tsagkogeorga et al., 2017).

For neglected tropical diseases, which disproportionately affect poorer people, it can be challenging to develop new medicines. To accelerate drug development in such cases, drug repurposing has been suggested whereby an already existing and approved medicine, or a well researched lead, is used to combat neglected tropical diseases (Ekins et al., 2011). Closantel, a veterinary anthelmintic has, for instance, been suggested for treatment of the human disease river blindness, caused by the filarial nematode Onchocerca volvulus (Gloeckner et al., 2010). As a first-pass bioinformatic identification of drug targets in four newly sequenced tapeworm genomes, OMA standalone was used to identify orthologs of known human drug targets (Tsai et al., 2013): Human genes targeted by drugs were retrieved from various databases, and their orthologs in tapeworms were inferred using OMA standalone. To identify targets likely to be essential across animals, orthologs with mice and nematodes were also identified: if both mice and nematode orthologs had knock-out phenotypes, we inferred that the orthologous group was essential across animals. Together with other indicators, such as gene expression data, we were able to rank every gene in these largely unexplored genomes for their suitability as a drug target, and associate lead compounds to them. As drugs could exhibit off-target effects on paralogs, the analysis focused on orthologs, which tend to be functionally more conserved (e.g. Altenhoff et al., 2012). The importance of investigating orthologs was illustrated by the drug Praziquantel, which is efficient against adult tapeworms, but not against the more dangerous larval form (Nogi et al., 2009). Praziquantel targets one particular voltage-gated calcium channel subunit. Using OMA standalone, we could identify the precise subunit ortholog in tapeworms and show that it is not expressed in the larval form—thereby providing a plausible explanation for the drug's low efficacy.

To conclude, orthology inference is a key step in integrating biological knowledge across multiple species. OMA standalone is a versatile orthology inference software with a proven track record. The software has been continuously improved and maintained over the past five years, undergoing 2 major and 25 minor (bug fixing) releases. We intend to keep developing and maintaining it. For support enquiries or bug reporting, we encourage users to use the biostars.org forum using the keyword "oma".

Material and Methods

Large-scale species phylogenetic reconstruction: Lophotrochozoa

We used transcriptome from seven Lophotrochozoa species published in (Egger et al., 2015): Mesodasys laticaudatus (Gastrotricha), Catenulida sp., Macrostomum ligano, Echinoplana celerrima, Microdalyellia schmidti, Monocelis sp. (Platyhelminthes) and Cerebratulus sp. (Nemertea). In addition,

12 sets of genomic and transcriptomic protein predictions from Saccoglossus kowalevskii, Brachionus plicatilis, Adineta ricciae, Schmidtea mediterranea, Lumbricus rubellus, Chaetopleura apiculata, Sepia officinalis, Mytilus californianus, Biomphalaria glabrata, Lymnaea stagnalis, Hydra magnipapillata and Amphimedon queenslandica were downloaded from the NCBI refseq repository (ftp.ncbi.nlm.nih.gov/refseq/). Redundant sequences with higher than 97% identity were removed by clustering with CD-HIT (Fu et al., 2012). Additionally, 11 precomputed proteomes for Homo sapiens, Strongylocentrotus purpuratus, Ciona intestinalis, Trichoplax adhaerens, Pristionchus pacificus, Caenorhabditis elegans, Drosophila melanogaster, Acyrthosiphon pisum, Capitella sp., Helobdella robusta and Lottia gigantea were downloaded from the OMA database website. The combined set of 30 non-redundant proteins sets contained 19 lophotrochozoans, four deuterostomes, four ecdysozoans, and proteomes from three non-bilaterian animals.

Quality assessment of sequencing reads was carried out with FastQC (Andrews and Others, 2010). Subsequent to this, it was determined, using PRINSEQ lite (Schmieder and Edwards, 2011), that the first 12 nucleotides should be trimmed off the 100bp reads. The assembly of the trimmed paired reads was done using Trinity v20130225 (Haas et al., 2013), with the flag '--min_kmer_cov 2', with default parameters.

In order to detect the presence of cross contaminations between the various libraries run on the same flow cell, we used the CroCo package (Simion et al., 2018). This identified any assembled transcripts with fewer than four read matches, which were subsequently discarded. Furthermore, this also discarded all transcripts in which the number of reads, from the intended species matching the transcript, was not at least five times greater than the number of matches to the transcript, from reads from any of the other potentially contaminating species.

For peptide predictions, all ORFs greater than 100aa were retained. For all peptide datasets, cd-hit was used to reduce redundancy by clustering sequences with a global sequence identity of greater than 95%.

For the HaMStR analysis, putative orthologs were determined for each species using HaMStR v13.2.6 (Ebersberger et al., 2009) using the Lophotrochozoa core ortholog set.

Orthologous groups were inferred by running BUSCO v1.22 (Simão et al., 2015) on the Metazoa dataset found at (https://busco.ezlab.org/v1/). We created orthologous groups made up of the protein sequences which BUSCO deemed to have had complete matches with their own highly conserved genes. At most one species containing multiple sequences was allowed per group. There was only a single occurrence of a group containing multiple sequences from a single species. In this case, we retained only the longest sequence.

The set of 30 proteomes were first filtered to remove low quality protein sequences using the OrthoMCL script "orthomclFilterFasta" (Chen et al., 2006). Low quality sequences were defined to be sequences that were shorter than 10 amino acids, contained more than 20% stop codons, and

contained more than 20% non-standard amino acids. An all versus all NCBI BLAST was then used with default parameters, in order to find the similarity score between sequences. Matches with an E-value < 10^{-6} were retained. Orthologs, in-paralogs and co-orthologs were then identified using the OrthoMCL script "OrthomclPairs" before clustering using MCL. An MCL inflation parameter of 2.2 was used in order to identify clusters. Each group was required to have at most one species containing multiple sequences. When more than one sequence from a single species was present, the longest sequence was selected to remain in the group, with the others removed.

Each orthologous group which contained a minimum of 15 protein sequences, of the 30 total, representing unique species were aligned using MUSCLE (Edgar, 2004), using default parameters. All spurious sequences, and poorly aligned regions of the multiple sequence alignments, were then removed using trimAl (Capella-Gutiérrez et al., 2009), using the -automated1 flag. Supermatrices were then constructed by concatenating all of the remaining alignments, with missing sequences treated as gaps. The final alignment was subsequently reduced to only contain sites in which more than 60% were occupied by amino acids.

Species trees were constructed using an LG+I model with 100 bootstrap replicates and a CAT+GTR+G4 model, with RAxMLv8.2.4 and PhyloBayes MPI v1.5a respectively. Convergence information is provided in Table 2.

Table 2: Convergence of the PhyloBayes runs

Method	Num Cycles	MaxDiff	MeanDiff
OMA	7,080	0.297691	0.00522266
HaMStR	2,731	0.297693	0.00972485
BUSCO	47,281	0.0957351	0.00435825
OrthoMCL	3,190	0.104071	0.0548237

Acknowledgements

Computations were performed on the University College London Computer Science cluster and at the Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. J.L is funded by EPSRC Centre for Doctoral Training studentship at UCL CoMPLEX (EP/F500351/1). MT acknowledges support by a Biotechnology and Biological Sciences Research Council grant (BB/H006966/1) and the European Research Council (ERC-2012-AdG 322790). C.D. acknowledges support by Swiss National Science Foundation grant 150654, UK BBSRC grant , and the Swiss State Secretariat for Education, Research and Innovation (SERI).

Competing interests

The authors declare no competing interests.

References

Afrasiabi C, Samad B, Dineen D, Meacham C, Sjölander K. 2013. The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res* **41**:W242–8.

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, Schreiber F, da Silva AS, Szklarczyk D, Train C-M, Bork P, Lecompte O, von Mering C, Xenarios I, Sjölander K, Jensen LJ, Martin MJ, Muffato M, Quest for Orthologs consortium, Gabaldón T, Lewis SE, Thomas PD, Sonnhammer E, Dessimoz C. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods* **13**:425–430.

Altenhoff AM, Dessimoz C. 2012. Inferring Orthology and Paralogy In: Anisimova M, editor. Evolutionary Genomics, Methods in Molecular Biology. Humana Press. pp. 259–279.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**:e1000262.

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* **8**:e53786.

Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, de Farias TM, Zile K, Stevenson C, Long J, Redestig H, Gonnet GH, Dessimoz C. 2017. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res*. doi:10.1093/nar/gkx1019

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, Gonnet GH, Dessimoz C. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res* **43**:D240–9.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* **8**:e1002514.

Andrews S, Others. 2010. FastQC: a quality control tool for high throughput sequence data.

Baguñà J, Riutort M. 2004. Molecular phylogeny of the Platyhelminthes. *Can J Zool* **82**:168–193.

Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C. 2011. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform* **12**:423–435.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

Carrigan MA, Uryasev O, Davis RP, Zhai L, Hurley TD, Benner SA. 2012. The natural history of class I primate alcohol dehydrogenases includes gene duplication, gene loss, and gene conversion. *PLoS One* **7**:e41175.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**:D363–8.

Dalquen D a., Dessimoz C. 2013. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biol Evol* **5**:1800–1806.

Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH. 2006. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res* **34**:3309–3316.

Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, Gonnet G. 2005. OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements In: McLysaght A, Huson DH, editors. RECOMB 2005 Workshop on Comparative Genomics. Springer-Verlag. pp. 61–72.

Dessimoz C, Gabaldón T, Roos DS, Sonnhammer ELL, Herrero J, Quest for Orthologs Consortium. 2012. Toward community standards in the quest for orthologs. *Bioinformatics* **28**:900–904.

Dikow RB, Frandsen PB, Turcatel M, Dikow T. 2017. Genomic and transcriptomic resources for assassin flies including the complete genome sequence of Proctacanthus coquilletti (Insecta: Diptera: Asilidae) and 16 representative transcriptomes. *PeerJ* **5**:e2951.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23**:533–539.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal Phylogeny and Its Evolutionary Implications*. *Annu Rev Ecol Evol Syst* **45**:371–395.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**:745–749.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**:157.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797.

Edgecombe GD, Giribet G, Dunn CW, Hejnol A, Kristensen RM, Neves RC, Rouse GW, Worsaae K, Sørensen MV. 2011. Higher-level metazoan relationships: recent progress and remaining questions. *Org Divers Evol* **11**:151–172.

Egger B, Lapraz F, Tomiczek B, Müller S, Dessimoz C, Girstmair J, Škunca N, Rawlinson KA, Cameron CB, Beli E, Todaro MA, Gammoudi M, Noreña C, Telford MJ. 2015. A Transcriptomic-Phylogenomic Analysis of the Evolutionary Relationships of Flatworms. *Curr Biol* **0**. doi:10.1016/j.cub.2015.03.034

Ekins S, Williams AJ, Krasowski MD, Freundlich JS. 2011. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov Today* **16**:298–310.

Fernández R, Giribet G. 2015. Unnoticed in the tropics: phylogenomic resolution of the poorly known arachnid order Ricinulei (Arachnida). *R Soc Open Sci* **2**:150065.

Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, Pérez-Porro AR, Edgecombe GD, Giribet G. 2014. Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. *Mol Biol Evol* **31**:1500–1513.

Forslund K, Pereira C, Capella-Gutierrez S, Sousa da Silva A, Altenhoff A, Huerta-Cepas J, Muffato M, Patricio M, Vandepoele K, Ebersberger I, Blake J, Fernández Breis JT, Quest for Orthologs Consortium, Boeckmann B, Gabaldón T, Sonnhammer E, Dessimoz C, Lewis S. 2017. Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*. doi:10.1093/bioinformatics/btx542

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14**:360–366.

Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, Hedin M, Kocot KM, Ledford JM, Bond JE. 2016. Spider phylogenomics: untangling the Spider Tree of Life. *PeerJ* **4**:e1719.

Giribet G, Edgecombe GD. 2017. Current Understanding of Ecdysozoa and its Internal Phylogenetic Relationships. *Integr Comp Biol* **57**:455–466.

Gloeckner C, Garner AL, Mersha F, Oksov Y, Tricoche N, Eubanks LM, Lustigman S, Kaufmann GF, Janda KD. 2010. Repositioning of an existing drug for the neglected tropical disease Onchocerciasis. *Proc Natl Acad Sci U S A* **107**:3424–3429.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**:1494–1512.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* **276**:4261–4270.

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res* **42**:D897–902.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**:D286–93.

Kocot KM. 2016. On 20 years of Lophotrochozoa. *Org Divers Evol* **16**:329–343.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, Halanych KM. 2011. Phylogenomics reveals deep molluscan relationships. *Nature* **477**:452–456.

Kocot KM, Struck TH, Merkel J, Waits DS, Todt C, Brannock PM, Weese DA, Cannon JT, Moroz LL, Lieb B, Halanych KM. 2017. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. *Syst Biol* **66**:256–282.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62**:611–615.

Laumer CE, Hejnol A, Giribet G. 2015. Nuclear genomic signals of the “microturbellarian” roots of platyhelminth evolutionary innovation. *Elife* **4**. doi:10.7554/eLife.05503

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**:2178–2189.

Linard B, Thompson JD, Poch O, Lecompte O. 2011. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics* **12**:11.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**:D183–D189.

Nogi T, Zhang D, Chan JD, Marchant JS. 2009. A novel biological activity of praziquantel requiring voltage-operated Ca2+ channel beta subunits: subversion of flatworm regenerative polarity. *PLoS Negl Trop Dis* **3**:e464.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**:D571–9.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**:4285–4288.

Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, Poustka AJ, Wallberg A, Peterson KJ, Telford MJ. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* **470**:255–258.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9**:e1000602.

Roth AC, Gonnet GH, Dessimoz C. 2008. Correction: Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**:518.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863–864.

Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform* **12**:485–488.

Sharma PP, Fernandez R, Santillan GR, Monod L. 2015. Phylogenomic resolution of scorpions reveals discordance with morphological phylogenetic signalINTEGRATIVE AND COMPARATIVE BIOLOGY. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA. pp. E165–E165.

Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol Biol Evol* **31**:2963–2984.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.

Simion P, Belkhir K, François C, Veyssier J, Rink JC, Manuel M, Philippe H, Telford MJ. 2018. A software tool "CroCo" detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol* **16**:28.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**:195–197.

Sonnhammer ELL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics* **30**:2993–2998.

Sonnhammer ELL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43**:D234–9.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Struck TH, Fisse F. 2008. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol Biol Evol* **25**:728–736.

Struck TH, Wey-Fabrizius AR, Golombek A, Hering L, Weigert A, Bleidorn C, Klebow S, Iakovenko N, Hausdorf B, Petersen M, Kück P, Herlyn H, Hankeln T. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. *Mol Biol Evol* **31**:1833–1849.

Szalkowski A, Ledergerber C, Krähenbühl P, Dessimoz C. 2008. SWPS3 - fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2. *BMC Res Notes* **1**:107.

Telford MJ, Budd GE, Philippe H. 2015. Phylogenomic Insights into Animal Evolution. *Curr Biol* **25**:R876–87.

Train C-M, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. 2017. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33**:i75–i82.

Tsagkogeorga G, Müller S, Dessimoz C, Rossiter SJ. 2017. Comparative genomics reveals contraction in olfactory receptor genes in bats. *Sci Rep* **7**:259.

Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, Aslett M, Beasley H, Bennett HM, Cai J, Camicia F, Clark R, Cucher M, De Silva N, Day T a., Deplazes P, Estrada K, Fernández C, Holland PWH, Hou J, Hu S, Huckvale T, Hung SS, Kamenetzky L, Keane J a., Kiss F, Koziol U, Lambert O, Liu K, Luo X, Luo Y, Macchiaroli N, Nichol S, Paps J, Parkinson J, Pouchkina-Stantcheva N, Riddiford N, Rosenzvit M, Salinas G, Wasmuth JD, Zamanian M, Zheng Y, Cai X, Soberón X, Olson PD, Laclette JP, Brehm K, Berriman M. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**:57–63.

Uchiyama I, Mihara M, Nishide H, Chiba H. 2012. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res* **41**:D631–5.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**:327–335.

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1618463114

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**:D744–D749.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2018.

Ensembl 2018. *Nucleic Acids Res* **46**:D754–D761.

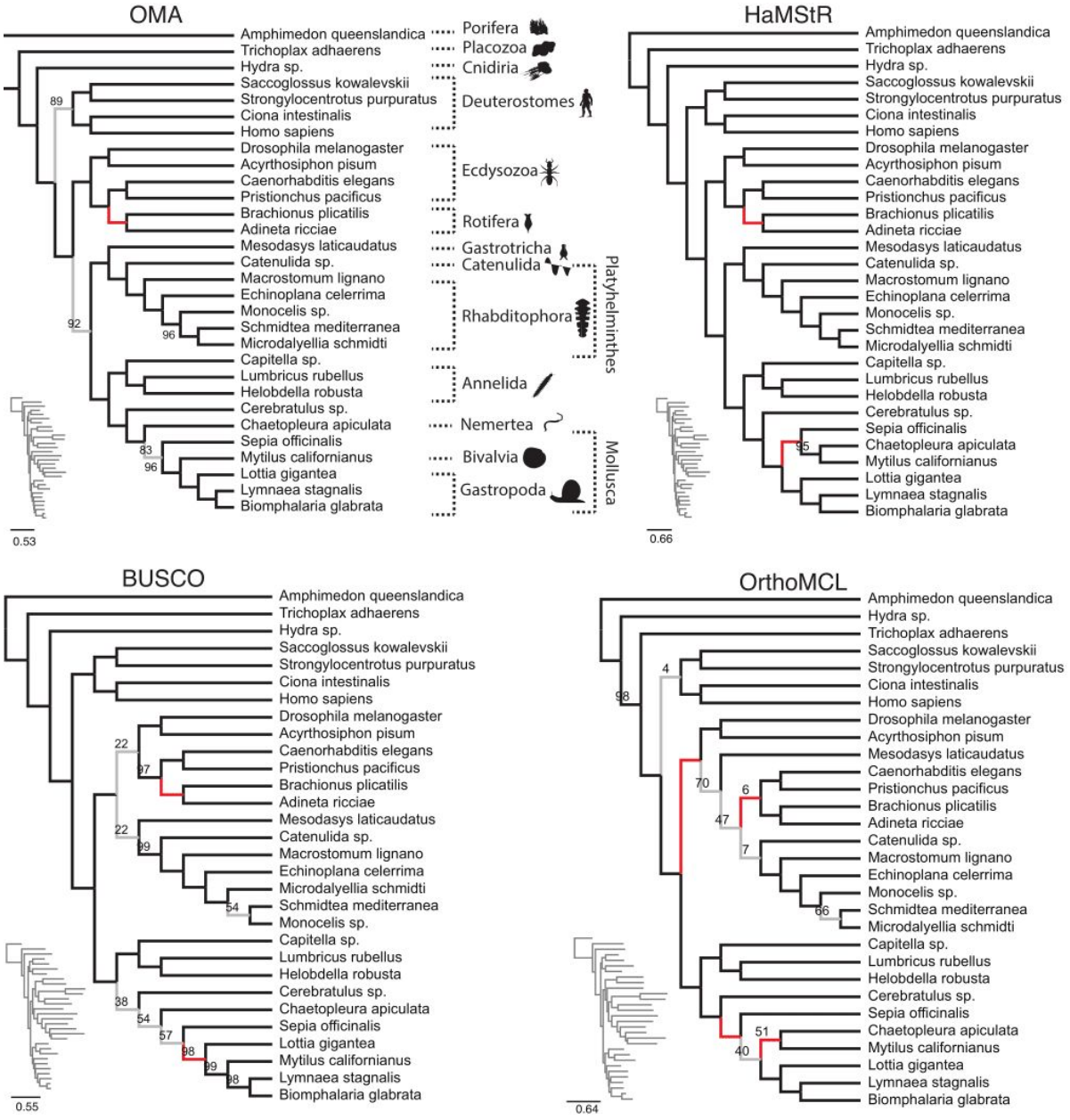


Fig. S1: Comparison of trees obtained using RAxML with the LG+I model for different datasets. OMA tree is in congruence with published results (see main text). Branches which are at odds with the literature are in red; else they are displayed in grey (posterior probability < 0.95) or else in black. Only posterior probabilities below 1 are displayed.

D. Bibliography

Aboobaker, A. A., Tomancak, P., Patel, N., Rubin, G. M., and Lai, E. C. *Drosophila* microRNAs exhibit diverse spatial expression patterns during embryonic development. *Proceedings of the National Academy of Sciences*, 102(50):18017–18022, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0508823102. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0508823102.

Achatz, J. G. and Martinez, P. The nervous system of *Isodiametra pulchra* (Acoela) with a discussion on the neuroanatomy of the Xenacoelomorpha and its evolutionary implications. *Frontiers in Zoology*, 9(1):27, 2012. ISSN 1742-9994. doi: 10.1186/1742-9994-9-27. URL http://www.frontiersinzoology.com/content/9/1/27/abstract.

Achatz, J. G., Chiodin, M., Salvenmoser, W., Tyler, S., and Martinez, P. The Acoela: On their kind and kinships, especially with nemertodermatids and xenoturbellids (Bilateria incertae sedis). *Organisms Diversity and Evolution*, 13(2):267–286, 2013. ISSN 14396092. doi: 10.1007/s13127-012-0112-4.

Altenhoff, A. M. and Dessimoz, C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, 5(1), 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000262.

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P., Schreiber, F., Da Silva, A. S., Szklarczyk, D., Train, C. M., Bork, P., Lecompte, O., Von Mering, C., Xenarios, I., Sjölander, K., Jensen, L. J., Martin, M. J., Muffato, M., Gabaldón, T., Lewis, S. E., Thomas, P. D., Sonnhammer, E., and Dessimoz, C. Standardized benchmarking in the quest for orthologs. *Nature Methods*, 13(5):425–430, 2016. ISSN 15487105. doi: 10.1038/nmeth.3830.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–10, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Auyeung, V. C., Ulitsky, I., McGeary, S. E., and Bartel, D. P. Beyond secondary structure: Primary-sequence determinants license Pri-miRNA hairpins for processing. *Cell*, 152 (4):844–858, 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.01.031. URL http://dx.doi.org/10.1016/j.cell.2013.01.031.

Balavoine, G. The early emergence of platyhelminths is contradicted by the agreement between 18S rRNA and Hox genes data. *C R Acad Sci III*, 320(1): 83–94, 1997. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=9099265.

Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116 (2):281–297, 2004. ISSN 00928674. doi: 10.1016/S0092-8674(04)00045-5.

Bartel, D. P. Metazoan MicroRNAs. *Cell*, 173(1):20–51, 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.03.006. URL http://dx.doi.org/10.1016/j.cell.2018.03.006.

Bentwich, I. Prediction and validation of microRNAs and their targets. *FEBS Letters*, 579(26):5904–5910, 2005. ISSN 00145793. doi: 10.1016/j.febslet.2005.09.040.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., and Bentwich, Z. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, 37(7): 766–770, 2005. ISSN 10614036. doi: 10.1038/ng1590.

Berezikov, E., Guryev, V., van De Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120(1):21–24, 2005. ISSN 00928674. doi: 10.1016/j.cell.2004.12.031.

Berkes, C. A. and Tapscott, S. J. MyoD and the transcriptional control of myogenesis. *Seminars in Cell and Developmental Biology*, 16(4-5):585–595, 2005. ISSN 10849521. doi: 10.1016/j.semcdb.2005.07.006.

Bortolomeazzi, M., Gaffo, E., and Bortoluzzi, S. A survey of software tools for microRNA discovery and characterization using RNA-seq. *Briefings in Bioinformatics*, (October): 1–13, 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx148.

Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. J., and Telford, M. J. *Xenoturbella* is a deuterostome that eats molluscs. *Nature*, 424(August):925–928, 2003. doi: 10.1038/nature01903.1.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E. S., Thorndyke, M. C., Nakano, H., Kohn, A. B., Heyland, A., Moroz, L. L., Copley, R. R., and Telford, M. J. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature*, 444 (7115):85–88, 2006. ISSN 0028-0836. doi: 10.1038/nature05241. URL `http://dx.doi.org/10.1038/nature05241`.

Bourlat, S. J., Nakano, H., Åkerman, M., Telford, M. J., Thorndyke, M. C., and Obst, M. Feeding ecology of *Xenoturbella bocki* (phylum Xenoturbellida) revealed by genetic barcoding. *Molecular Ecology Resources*, 8(1):18–22, 2008. ISSN 1755098X. doi: 10.1111/j.1471-8286.2007.01959.x.

Bourlat, S. J., Rota-Stabelli, O., Lanfear, R., and Telford, M. J. The mitochondrial genome structure of *Xenoturbella bocki* (phylum Xenoturbellida) is ancestral within the deuterostomes. *BMC Evolutionary Biology*, 9:107, 2009. ISSN 1471-2148. doi: 10.1186/1471-2148-9-107.

Brooke, N. M., Garcia-Fernàndez, J., and Holland, P. W. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679):920–922, 1998. ISSN 00280836. doi: 10.1038/31933.

Brown, N. P., Leroy, C., and Sander, C. MView: a web-compatible database search or multiple alignment viewer. *Oxford University Press*, 14(4):380–381, 1998. ISSN 1367-4803. doi: 10.1093/bioinformatics/14.4.380.

Buchfink, B., Xie, C., and Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3176. URL `http://www.nature.com/doifinder/10.1038/`

nmeth.3176{%}5Cnhttp://dx.doi.org/10.1038/nmeth.3176{%}5Cnhttp://www.nature.com/doifinder/10.1038/nmeth.3176{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25402007.

Campbell, L. I., Rota-Stabelli, O., Edgecombe, G. D., Marchioro, T., Longhorn, S. J., Telford, M. J., Philippe, H., Rebecchi, L., Peterson, K. J., and Pisani, D. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proceedings of the National Academy of Sciences*, 108(38):15920–15924, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1105499108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1105499108.

Campos, A., Cummings, M. P., Reyes, J. L., and Laclette, J. P. Phylogenetic relationships of Platyhelminthes based on 18S ribosomal gene sequences. *Mol. Phylogenet. Evol.*, 10(1):1–10, 1998.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, 530(7588):89–93, 2016. ISSN 0028-0836. doi: 10.1038/nature16520. URL http://www.nature.com/doifinder/10.1038/nature16520.

Carranza, S., Baguñà, J., and Riutort, M. Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences. *Mol Biol Evol*, 14(5):485–497, 1997. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025785. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed{&}cmd=Retrieve{&}dopt=AbstractPlus{&}list{_}uids=9159926.

Castresana, J., Feldmaier-Fuchs, G., Yokobori, S.-i., Satoh, N., and Pääbo, S. The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics*, 150(3):1115–1123, 1998. ISSN 00166731.

Cook, C. E., Jiménez, E., Akam, M., and Saló, E. The Hox gene complement of acoel flatworms, a basal bilaterian clade. *Evolution and Development*, 6(3):154–163, 2004. ISSN 1520541X. doi: 10.1111/j.1525-142X.2004.04020.x.

Dalquen, D. A. and Dessimoz, C. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biology and Evolution*, 5 (10):1800–1806, 2013. ISSN 17596653. doi: 10.1093/gbe/evt132.

de Mendoza, A. and Ruiz-Trillo, I. The mysterious evolutionary origin for the GNE gene and the root of Bilateria. *Molecular Biology and Evolution*, 28(11):2987–2991, 2011. ISSN 07374038. doi: 10.1093/molbev/msr142.

de Rosa, R., Grenler, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., Carroll, S. B., and Balavoine, G. Hox genes in brachiopods and priapulids and protostome evolution. *Nature*, 399(6738):772–776, 1999. ISSN 00280836. doi: 10.1038/21631.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–9, 2008. ISSN 1476-4687. doi: 10.1038/nature06614. URL http://www.ncbi.nlm.nih.gov/pubmed/18322464.

Ehlers, U. and Sopott-Ehlers, B. Ultrastructure of the subepidermal musculature of *Xenoturbella bocki*, the adelphotaxon of the Bilateria. *Zoomorphology*, 117:71–79, 1997. ISSN 0720-213X. doi: 10.1007/s004350050032.

Emms, D. M. and Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):157, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0721-2. URL http://dx.doi.org/10.1186/s13059-015-0721-2.

Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1344.

Franzén, A. On spermiogenesis, morphology of the spermatozoon, and biology of fertilization among invertebrates. *Zool Bidrag, Uppsala*, 31:355–482, 1956.

Franzén, Å. and Afzelius, B. A. The ciliated epidermis of *Xenoturbella bocki* (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. *Zoologica Scripta*, 16(1):9–17, 1987. ISSN 14636409. doi: 10.1111/j.1463-6409.1987.tb00046.x.

Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W., and Rajewsky, N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2012. ISSN 03051048. doi: 10.1093/nar/gkr688.

Fritzsch, G., Böhme, M. U., Thorndyke, M. C., Nakano, H., Israelsson, O., Stach, T., Schlegel, M., Hankeln, T., and Stadler, P. F. PCR survey of *Xenoturbella bocki* Hox genes. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 310(3):278–284, 2008. ISSN 15525007. doi: 10.1002/jez.b.21208.

Fromm, B., Worren, M. M., Hahn, C., Hovig, E., and Bachmann, L. Substantial loss of conserved and gain of novel MicroRNA families in flatworms. *Molecular Biology and Evolution*, 30(12):2619–2628, 2013. ISSN 07374038. doi: 10.1093/molbev/mst155.

Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., Newcomb, J. M., Sempere, L. F., Flatmark, K., Hovig, E., and Peterson, K. J. A Uniform System For The Annotation Of Human microRNA Genes And The Evolution Of The Human microRNAome. *Annu Rev Genet.*, 23(49):213–242, 2015. doi: 10.1146/annurev-genet-120213-092023. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4743252/pdf/nihms754425.pdf`.

Ha, M. and Kim, V. N. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8):509–524, 2014. ISSN 14710080. doi: 10.1038/nrm3838. URL `http://dx.doi.org/10.1038/nrm3838`.

Han, J., Morris, S. C., Ou, Q., Shu, D., and Huang, H. Meiofaunal deuterostomes from the basal Cambrian of Shaanxi (China). *Nature*, 542(7640):228–231, 2017. ISSN 14764687. doi: 10.1038/nature21072. URL `http://dx.doi.org/10.1038/nature21072`.

Haszprunar, G. Review of data for a morphological look on Xenacoelomorpha (Bilateria incertae sedis). *Organisms Diversity and Evolution*, 16(2):363–389, 2016. ISSN 16181077. doi: 10.1007/s13127-015-0249-z.

Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. J., and Peterson, K. J. MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):2946–50, 2008. ISSN 1091-6490. doi: 10.1073/pnas.0712259105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2268565{&}tool=pmcentrez{&}rendertype=abstract.

Heimberg, A. M., Cowper-Sal·lari, R., Sémon, M., Donoghue, P. C. J., and Peterson, K. J. MicroRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proceedings of the National Academy of Sciences*, 107(45):19379–19383, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1010350107. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1010350107.

Hejnol, A. and Martindale, M. Q. Acoel development supports a simple planula-like urbilaterian. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1496):1493–501, 2008. ISSN 0962-8436. doi: 10.1098/rstb.2007.2239. URL http://www.ncbi.nlm.nih.gov/pubmed/18192185.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., and Dunn, C. W. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B*, 276(September):4261–4270, 2009. ISSN 1471-2954. doi: 10.1098/rspb.2009.0896. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817096{&}tool=pmcentrez{&}rendertype=abstract.

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I. L., and Stadler, P. F. The expansion of the metazoan microRNA repertoire. *BMC genomics*, 7:25, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-25. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1388199{&}tool=pmcentrez{&}rendertype=abstract.

Israelsson, O. ...and Molluscan Embryogenesis. *Nature*, 390(6655):32, 1997. ISSN 00280836. doi: 10.1038/36246.

Israelsson, O. New light on the enigmatic *Xenoturbella* (phylum uncertain): ontogeny and phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, 266(October 1998):835, 1999. ISSN 0962-8452. doi: 10.1098/rspb.1999.0713.

Jiménez-Guri, E., Paps, J., García-Fernàndez, J., and Saló, E. Hox and ParaHox genes in Nemertodermatida, a basal bilaterian clade. *International Journal of Developmental Biology*, 50(8):675–679, 2006. ISSN 02146282. doi: 10.1387/ijdb.062167ej.

Jondelius, U., Ruiz-Trillo, I., Baguñà, J., and Riutort, M. The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. *Zoologica Scripta*, 31(2): 201–215, 2002. ISSN 03003256. doi: 10.1046/j.1463-6409.2002.00090.x.

Jondelius, U., Wallberg, A., Hooge, M., and Raikova, O. I. How the worm got its pharynx: Phylogeny, classification and bayesian assessment of character evolution in Acoela. *Systematic Biology*, 60(6):845–871, 2011. ISSN 10635157. doi: 10.1093/sysbio/syr073.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu031.

Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. MicroRNAs and Their Regulatory Roles in Plants. *Annual Review of Plant Biology*, 57(1):19–53, 2006. ISSN 1543-5008. doi: 10.1146/annurev.arplant.57.032905.105218. URL `http://www.annualreviews.org/doi/10.1146/annurev.arplant.57.032905.105218`.

Kadri, S., Hinman, V., and Benos, P. V. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, 10(Suppl 1):S35, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S1-S35. URL `http://www.biomedcentral.com/1471-2105/10/S1/S35`.

Katayama, T., Nishioka, M., and Yamamoto, M. Phylogenetic relationships among turbellarian orders inferred from 18S rDNA sequences. *Zoological Science*, 13(5): 747–756, 1996. ISSN 0289-0003. doi: 10.2108/zsj.13.747.

Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39:309–338, 2005. ISSN 0066-4197. doi: 10.1146/annurev.genet.39.073003.114725.

Kozomara, A. and Griffiths-Jones, S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):68–73, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1181.

Krämer-Eis, A., Feretti, L., Schiffer, P. H., Heger, P., and Wiehe, T. The common developmental genetic toolkit of bilaterian crown clades after a billion years of divergence, 2016.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., and Koonin, E. V. Computational methods for Gene Orthology inference. *Briefings in Bioinformatics*, 12(5):379–391, 2011. ISSN 14675463. doi: 10.1093/bib/bbr030.

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biology*, 4(7):R42, 2003. ISSN 1465-6906. doi: 10.1186/gb-2003-4-7-r42. URL http://genomebiology.com/2003/4/7/R42.

Lau, N. C., Lim, L. P., Weistein, E. G., and Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294(October):858–862, 2001.

Lee, R. C. and Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 294(OCTOBER):862–864, 2001.

Lee, R. C., Feinbaum, R. L., and Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993. ISSN 00928674. doi: 10.1016/0092-8674(93)90529-Y.

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V. N. MicroRNA maturation: stepwisee processing and subcellular localization. *The EMBO Journal*, 21 (17):4663–4670, 2002. ISSN 0261-4189. doi: 10.1093/emboj/cdf476. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=

pubmed{&}id=12198168{&}retmode=ref{&}cmd=prlinks{%}5Cnpapers3://publication/uuid/2F4193DD-DD3A-4F58-97A2-C99CC338E6F4.

Lemons, D. and McGinnis, W. Genomic evolution of Hox gene clusters. *Science*, 313 (September):1918–1922, 2006. ISSN 8657188206636. doi: 10.1126/science.1132040.

Li, L., Stoeckert, C. J. J., and Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189, 2003. ISSN 1088-9051. doi: 10.1101/gr.1224503.candidates. URL http://genome.cshlp.org/cgi/content/full/13/9/2178.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. The microRNAs of *C. elegans*. *Genes & Development*, pages 991–1008, 2003. doi: 10.1101/gad.1074403.regulating.

Linard, B., Thompson, J. D., Poch, O., and Lecompte, O. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12:11, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-11. URL papers3://publication/doi/10.1186/1471-2105-12-11.

Ling, H., Fabbri, M., and Calin, G. A. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews Drug Discovery*, 12(11):847–865, 2013. ISSN 14741776. doi: 10.1038/nrd4140. URL http://dx.doi.org/10.1038/nrd4140.

Littlewood, D. T. J., Rohde, K., Bray, R. A., and Herniou, E. A. Phylogeny of the Platyhelminthes and the evolution of parasitism. *Biological Journal of the Linnean Society*, 68(1-2):257–287, 1999. ISSN 00244066. doi: 10.1006/bijl.1999.0341. URL http://www.sciencedirect.com/science/article/pii/S0024406699903413.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26. URL http://almob.biomedcentral.com/articles/10.1186/1748-7188-6-26.

Lundin, K. Degenerating epidermal cells in *Xenoturbella bocki* (phylum uncertain), Nemertodermatida and Acoela (Platyhelminthes). *Belgian Journal of Zoology*, 131 (Suppl. 1):153–157, 2001.

Lundin, K. and Hendelberg, J. Is the sperm type of the Nemertodermatida close to that of the ancestral Platyhelminthes? *Hydrobiologia*, 383:197–205, 1998. ISSN 00188158. doi: 10.1023/A:1003439512957.

Martin, M. W., Grazhdankin, D. V., Bowring, S. A., Evans, D. A. D., Fedonkin, M. A., and Kirschvink, J. L. Age of Neoproterozoic bilatarian body and trace fossils, White Sea, Russia: Implications for metazoan evolution. *Science*, 288(5467):841–845, 2000.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(June):1344–1349, 2008.

Nakano, H., Miyazawa, H., Maeno, A., Shiroishi, T., Kakui, K., Koyanagi, R., Kanda, M., Satoh, N., Omori, A., and Kohtsuka, H. A new species of *Xenoturbella* from the western Pacific Ocean and the evolution of *Xenoturbella*. *BMC Evolutionary Biology*, 17(1):1–11, 2017. ISSN 14712148. doi: 10.1186/s12862-017-1080-2.

Nielsen, C. *Animal Evolution*. 1995.

Norén, M. and Jondelius, U. *Xenoturbella*'s molluscan relatives. *Nature*, 390:31–32, 1997. ISSN 00280836. doi: 10.1038/36242.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. L. Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33(DATABASE ISS.):476–480, 2005. ISSN 03051048. doi: 10.1093/nar/gki107.

Obst, M., Nakano, H., Bourlat, S. J., Thorndyke, M. C., Telford, M. J., Nyengaard, J. R., and Funch, P. Spermatozoon ultrastructure of *Xenoturbella bocki* (Westblad 1949). *Acta Zoologica*, 92(2):109–115, 2011. ISSN 00017272. doi: 10.1111/j.1463-6395.2010.00496.x.

Okamura, K., Chung, W. J., and Lai, E. C. The long and short of inverted repeat genes in animals: MicroRNAs, mirtrons and hairpin RNAs. *Cell Cycle*, 7(18):2840–2845, 2008. ISSN 15514005. doi: 10.4161/cc.7.18.6734.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1189.

Paps, J. and Holland, P. W. H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nature Communications*, 9(1730):1–8, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04136-5. URL `http://dx.doi.org/10.1038/s41467-018-04136-5`.

Pardos, F. Fine Structure and Function of Pharynx Cilia in *Glossobalanus minutus* Kowalewsky (Enteropneusta). *Acta Zoologica*, 69(1):1–12, 1988.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000. ISSN 0028-0836. doi: 10.1038/35040556.

Pasquinelli, A. E., McCoy, A., Jiménez, E., Saló, E., Ruvkun, G., Martindale, M. Q., and Baguñà, J. Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: A role in life history evolution? *Evolution and Development*, 5(4):372–378, 2003. ISSN 1520541X. doi: 10.1046/j.1525-142X.2003.03044.x.

Paul, P., Chakraborty, A., Sarkar, D., Langthasa, M., Rahman, M., Bari, M., Singha, R. K. S., Malakar, A. K., and Chakraborty, S. Interplay between miRNAs and human

diseases. *Journal of Cellular Physiology*, 233(3):2007–2018, 2017. doi: 10.1002/jcp.25854.

Pedersen, K. J. and Pedersen, L. R. Ultrastructural Observations on the Epidermis of *Xenoturbella bocki* Westblad, 1949; With a Discussion of Epidermal Cytoplasmic Filament Systems of Invertebrates. *Acta Zoologica*, 69(4):231–246, 1988. ISSN 14636395. doi: 10.1111/j.1463-6395.1988.tb00920.x.

Perseke, M., Hankeln, T., Weich, B., Fritzsch, G., Stadler, P. F., Israelsson, O., Bernhard, D., and Schlegel, M. The mitochondrial DNA of *Xenoturbella bocki*: Genomic architecture and phylogenetic analysis. *Theory in Biosciences*, 126(1):35–42, 2007. ISSN 14317613. doi: 10.1007/s12064-007-0007-7.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., and Telford, M. J. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*, 470:255–258, 2011. ISSN 0028-0836. doi: 10.1038/nature09676.

Pinzón, N., Li, B., Martinez, L., Sergeeva, A., Presumey, J., Apparailly, F., and Seitz, H. MicroRNA target prediction programs predict many false positives. *Genome research*, 27(2):234–245, 2017. ISSN 1549-5469. doi: 10.1101/gr.205146.116. URL http://www.ncbi.nlm.nih.gov/pubmed/28148562http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5287229.

Prochnik, S. E., Rokhsar, D. S., and Aboobaker, A. A. Evidence for a microRNA expansion in the bilaterian ancestor. *Development Genes and Evolution*, 217(1):73–77, 2007. ISSN 0949944X. doi: 10.1007/s00427-006-0116-1.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., Jurka, J., Genikhovich, G., Grigoriev, I. V., Lucas, S. M., Steele, R. E., Finnerty, J. R., Technau, U., Martindale, M. Q., and Rokhsar, D. S. Sea anemone genome reveals the gene repertoire and genomic organization of the eumetazoan ancestor. *Science*, 317(5834):86–94, 2007. doi: 10.1126/science.1139158.

Reisinger, E. Was ist *Xenoturbella*? *Zeitschrift für wissenschaftliche Zoologie*, 164: 188–198, 1960.

Robertson, H. E. *Molecular approaches for studying the evolution of the Xenacoelomorpha*. PhD thesis, University College London, 2017.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10 A):1902–1910, 2004. ISSN 10889051. doi: 10.1101/gr.2722704.

Rohde, K., Watson, N., and Cannon, L. Ultrastructure of epidermal cilia of *Pseudactinoposthia* sp. (Platyhelminthes, Acoela); implications for the phylogenetic status of the Xenoturbellida and Acoelomorpha. *Journal of submicroscopic cytology*, 20(4): 759–767, 1988.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J., Pisani, D., Philippe, H., and Telford, M. J. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B: Biological Sciences*, 278 (1703):298–306, 2011. ISSN 0962-8452. doi: 10.1098/rspb.2010.0590. URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2010.0590.

Rouse, G. W., Wilson, N. G., Carvajal, J. I., and Vrijenhoek, R. C. New deep-sea species of *Xenoturbella* and the position of Xenacoelomorpha. *Nature*, 530(7588):94–97, 2016. ISSN 0028-0836. doi: 10.1038/nature16545. URL http://www.nature.com/doifinder/10.1038/nature16545.

Ruiz-Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A., and Baguñà, J. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science*, 283(5409):1919–1923, 1999. ISSN 00368075. doi: 10.1126/science.283.5409.1919. URL http://www.sciencemag.org/cgi/doi/10.1126/science.283.5409.1919.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Baguñà, J., and Riutort, M. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17):11246–11251, 2002. ISSN 00278424 (ISSN). doi: 10.1073/

pnas.172390199. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-0037143719{&}partnerID=40{&}md5=82e676d14de7371a645d6f9203391ef4.

Sempere, L. F., Cole, C. N., McPeek, M. A., and Peterson, K. J. The Phylogenetic Distribution of Metazoan microRNAs: Insights into Evolutionary Complexity and Constraint. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 306(6):575–588, 2006. ISSN 15525007. doi: 10.1002/jez.b.

Sempere, L. F., Martinez, P., Cole, C., Baguñà, J., and Peterson, K. J. Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. *Evolution and Development*, 9(5):409–415, 2007. ISSN 1520541X. doi: 10.1111/j.1525-142X.2007.00180.x.

Shabalina, S. A. and Koonin, E. V. Origins and evolution of eukaryotic RNA interference. *Trends in Ecology and Evolution*, 23(10):578–587, 2008. ISSN 01695347. doi: 10.1016/j.tree.2008.06.005.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.75. URL http://msb.embopress.org/content/7/1/539.abstract.

Simakov, O., Marlétaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., Boore, J. L., Grigoriev, I. V., Lindberg, D. R., Seaver, E. C., Weisblat, D. A., Putnam, N. H., and Rokhsar, D. S. Insights into bilaterian evolution from three spiralian genomes. *Nature*, 493 (7433):526–531, 2013. ISSN 0028-0836. doi: 10.1038/nature11696. URL http://www.nature.com.proxyiub.uits.iu.edu/nature/journal/v493/n7433/abs/nature11696.html{%}5Cnhttp://www.nature.com.proxyiub.uits.iu.edu/nature/journal/v493/n7433/pdf/nature11696.pdf.

Simakov, O., Kawashima, T., Marlétaz, F., Jenkins, J., Koyanagi, R., Mitros, T., Hisata, K., Bredeson, J., Shoguchi, E., Gyoja, F., Yue, J.-X., Chen, Y.-C., Freeman, R. M.,

Sasaki, A., Hikosaka-Katayama, T., Sato, A., Fujie, M., Baughman, K. W., Levine, J., Gonzalez, P., Cameron, C., Fritzenwanker, J. H., Pani, A. M., Goto, H., Kanda, M., Arakaki, N., Yamasaki, S., Qu, J., Cree, A., Ding, Y., Dinh, H. H., Dugan, S., Holder, M., Jhangiani, S. N., Kovar, C. L., Lee, S. L., Lewis, L. R., Morton, D., Nazareth, L. V., Okwuonu, G., Santibanez, J., Chen, R., Richards, S., Muzny, D. M., Gillis, A., Peshkin, L., Wu, M., Humphreys, T., Su, Y.-H., Putnam, N. H., Schmutz, J., Fujiyama, A., Yu, J.-K., Tagawa, K., Worley, K. C., Gibbs, R. A., Kirschner, M. W., Lowe, C. J., Satoh, N., Rokhsar, D. S., and Gerhart, J. Hemichordate genomes and deuterostome origins. *Nature*, 527(7579):459–465, 2015. ISSN 0028-0836. doi: 10.1038/nature16150. URL http://www.nature.com/doifinder/10.1038/nature16150.

Srivastava, M., Mazza-Curll, K. L., Van Wolfswinkel, J. C., and Reddien, P. W. Whole-body acoel regeneration is controlled by Wnt and Bmp-Admp signaling. *Current Biology*, 24(10):1107–1113, 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.03.042. URL http://dx.doi.org/10.1016/j.cub.2014.03.042.

Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu033.

Sukumaran, J. and Holder, M. T. DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq228.

Tautz, D. and Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nature reviews. Genetics*, 12(10):692–702, 2011. ISSN 1471-0064. doi: 10.1038/nrg3053. URL http://www.ncbi.nlm.nih.gov/pubmed/21878963.

Telford, M. J. and Copley, R. R. Zoology: War of the Worms. *Current Biology*, 26 (8):R335–R337, 2016. ISSN 09609822. doi: 10.1016/j.cub.2016.03.015. URL http://dx.doi.org/10.1016/j.cub.2016.03.015.

Telford, M. J., Herniou, E. A., Russell, R. B., and Littlewood, D. T. J. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proceedings of the National Academy of Sciences of the United S.tates of America*,

97(21):11359–11364, 2000. ISSN 0027-8424. doi: 10.1073/pnas.97.21.11359. URL http://www.hubmed.org/display.cgi?uids=11027335.

Telford, M. J., Lockyer, A. E., Cartwright-Finch, C., and Littlewood, D. T. J. Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. *Proceedings. Biological sciences / The Royal Society*, 270(1519):1077–83, 2003. ISSN 0962-8452. doi: 10.1098/rspb.2003.2342. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1691347{&}tool=pmcentrez{&}rendertype=abstract.

Thomson, R. C., Plachetzki, D. C., Mahler, D. L., and Moore, B. R. A critical appraisal of the use of microRNA data in phylogenetics. *Proceedings of the National Academy of Sciences*, 111(35):E3659–E3668, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1407207111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1407207111.

Tyler, S. Distinctive features of cilia in metazoons and their significance for systematics. *Tissue & Cell*, 11(3):385–400, 1979.

van Dongen, S. Graph clustering by flow simulation. *Graph stimulation by flow clustering*, PhD thesis:University of Utrecht, 2000. ISSN 15740137. doi: 10.1016/j.cosrev.2007.05.001.

Westblad, E. *Xenoturbella bocki* ng, n. sp., a peculiar, primitive turbellarian type, 1949.

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., and Peterson, K. J. The deep evolution of metazoan microRNAs. *Evolution and Development*, 11(1):50–68, 2009. ISSN 1520541X. doi: 10.1111/j.1525-142X.2008.00302.x.

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H. R., Kauppinen, S., and Plasterk, R. H. A. MicroRNA expression in zebrafish embryonic development Ant Nestmate and Non-Nestmate discrimination by a chemosensory Sensillum. *Science*, 309(July):310–311, 2005.

Wightman, B., Ha, I., and Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5):855–862, 1993. ISSN 00928674. doi: 10.1016/0092-8674(93)90530-4.

Winter, J., Jung, S., Keller, S., Gregory, R. I., and Diederichs, S. Many roads to maturity: MicroRNA biogenesis pathways and their regulation. *Nature Cell Biology*, 11(3):228–234, 2009. ISSN 14657392. doi: 10.1038/ncb0309-228.

Wolf, Y. I. and Koonin, E. V. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biology and Evolution*, 4(12):1286–1294, 2012. ISSN 17596653. doi: 10.1093/gbe/evs100.

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. Conservation and divergence of plant microRNA genes. *Plant Journal*, 46(2):243–259, 2006. ISSN 09607412. doi: 10.1111/j.1365-313X.2006.02697.x.